DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

SELECTING THE BEST LINEAR REGRESSION MODEL: A CLASSICAL APPROACH

Donald Lien and Quang H. Vuong
California Institute of Technology

SOCIAL SCIENCE WORKING PAPER 606

March 1986

SELECTING THE BEST LINEAR REGRESSION MODEL: A CLASSICAL APPROACH

Donald Lien and Quang H. Vuong
Division of Humanities and Social Sciences
California Institute of Technology
Pasadena, CA 91125

ABSTRACT

In this paper, we apply the model selection approach based on Likelihood Ratio (LR) tests developed in Vuong (1985) to the problem of choosing between two normal linear regression models which are not nested in each other. First we compare our model selection procedure to other model selection criteria. Then we explicitly derive the procedure when the competing linear models are non-nested and neither one is correctly specified. Some simplifications are seen to arise when both models are contained in a larger correctly specified linear regression model, or when at least one competing linear model is correctly specified. A comparison of our model selection tests and previous non-nested hypothesis tests concludes the paper.

SELECTING THE BEST LINEAR REGRESSION MODEL: A CLASSICAL APPROACH

Donald Lien and Quang H. Vuong
California Institute of Technology

1. INTRODUCTION

In this paper, we apply the model selection approach developed in Vuong (1985) to the classical problem of choosing between two linear regression models. That this problem is important to applied econometricians results from the fact that one does not in general have a unique econometric model either because one wants to compare many theories or because a theory does not provide a unique functional form.

The problem of selecting the "best" subset of variables in a linear regression context has long been of special interest to theoretical and applied statisticians. The numerous papers that were generated by this classical problem have recently been surveyed by Gaver and Geisel (1974), Hocking (1976), and Lindley (1968) among others. Various solutions dealing with different aspects of this problem were proposed.

A first general approach is to set the problem in a decision framework. The natural solution is then the Bayesian solution which relies on Jeffrey's posterior probability criterion (see in particular Zellner (1971)). This solution will not be discussed in this paper. Other solutions in a decision theoretic framework but which are not always justified in a Bayesian setting are based on the construction of model selection criteria. Most of these widely used criteria will

be discussed and compared to our solution later in this paper.

A second general approach is to adopt the classical hypothesis testing framework. In this context, two solutions are in general accepted. The first solution derives from the work of Cox (1960, 1961) on testing non-nested hypotheses. Starting with Pesaran (1974), this solution has recently attracted a lot of attention from theoretical econometricians. The second solution consists in nesting the competing models in a larger model and to test that the additional parameters are equal to some particular values (Atkinson (1970)).

Within the classical hypothesis framework, there is however a third solution which has not been widely recognized and which dates back to Hotelling (1940). It consists in discriminating between the competing models by testing the hypothesis that the models are "equivalent" under some appropriate definition. Recent works along this line are White and Olson (1979) where the mean squared error of prediction is used, and Vuong (1985) where the Kullback-Leibler (1951) criterion is used. The advantage of this discriminating approach is that, unlike other classical solutions, the competing models are treated symmetrically.

The purpose of this paper is to develop this discriminating solution for the case where the competing models are normal linear regressions. Since neither model may be correctly specified, by necessity, this paper is mainly concerned with asymptotic results. The paper is organized as follows. In Section 2, we discuss two theoretical model selection criteria, i.e., mean squared error (MSE)

of prediction and Kullback-Leibler information criterion (KLIC). They represent two different distance measures between two probability distributions. Upon applying the two criteria to normal linear regression models, we show that they lead to the same comparison. In Section 3, we show that most of the model selection criteria in the literature are either consistent estimates of MSE or consistent estimates of KLIC. Based on this remark, a short survey is provided.

We then turn to the model selection approach based on Likelihood Ratio (LR) tests that is developed in Vuong (1985). Specifically, we characterize this procedure when the competing linear regression models are non-nested and neither one is correctly specified. The complicated results are presented in Section 4. Some simplifications are seen to arise when both models are contained in a larger linear regression model which is correctly specified or when at least one model is correctly specified. These results are discussed in Section 5 and Section 6, respectively. A comparison of our model selection tests and previous non-nested hypothesis tests concludes the paper. All the proofs are collected in the Appendix.

2. TWO THEORETICAL MODEL SELECTION CRITERIA

In this section we consider two important theoretical model selection criteria and we apply them to the normal linear regression model. We shall argue in the next section that most of the current model selection criteria can be thought of as estimates of either one of these two theoretical criteria. Unlike previous studies on model

selection that assume fixed explanatory variables or fixed in repeated sample, we shall assume that our explanatory variables are random, an assumption that is justified with economic data. Let $(y_t, \underline{x}_t')'$ be the t-th observation on the $(1 + \ell)$-dimensional random vector $(y, \underline{x}')'$ defined on an Euclidean measurable space. For simplicity, we adopt Vuong (1983) framework and assume:

Assumption A1: The random vectors $(y_t, \underline{x}_t')'$, $t = 1,2,\ldots$ are independent and identically distributed with common true cumulative distribution function H^o.[1]

In econometric modelling, we are interested in the true conditional distribution of y given $\underline{x}$. Let $H^o_{y|\underline{x}}(\cdot|\cdot)$ denote such a distribution. To estimate $H^o_{y|\underline{x}}(\cdot|\cdot)$, we specify parametric conditional models for y given $\underline{x}$, i.e., parametric families of conditional distributions for y given $\underline{x}$, $F_\theta = \{F_{y|\underline{x}}(\cdot|\cdot;\theta), \theta \in \Theta\}$. In this paper, we shall consider linear regression models with normal errors. Each linear regression model will be associated with a subset of the "exogenous" variables $\underline{x}$. Specifically, let $\underline{x}_s$ be a ℓ_s-subset of $\underline{x}$. Then, the normal linear regression model for y given $\underline{x}$ with explanatory variables $\underline{x}_s$ is formally defined as:

$$M_s \equiv \{N(\lambda_c + \underline{x}_s'\underline{\lambda}_s, \sigma_s^2);\ \theta_s \equiv (\lambda_c, \underline{\lambda}_s', \sigma_s^2)' \in \mathbb{R}^{\ell_s+1} \times \mathbb{R}_+\} \qquad (2.1)$$

where $N(\mu,\sigma^2)$ denotes the univariate normal distribution with mean μ and variance σ^2. Thus the θ-conditional distribution for y given $\underline{x}$ in M_S specifies that $E_\theta(y|\underline{x}) = \lambda_c + \underline{x}_S'\underline{\lambda}_S$ and $Var_\theta(y|\underline{x}) = \sigma_S^2$.

To evaluate the adequacy of a specified conditional model for y given $\underline{x}$, it is first necessary to define a measure of distance between the true conditional distribution $H^o_{y|\underline{x}}$ and a given conditional distribution $F_{y|\underline{x}}(\theta)$. Two measures of distance are generally accepted.

The first measure is based on the mean squared error (MSE) of prediction. Let $E_\theta(y|\underline{x})$ be the conditional expectation of y given $\underline{x}$ for the conditional distribution $F_{y|\underline{x}}(\theta)$, i.e.,

$$E_\theta(y|\underline{x}) \equiv \int y \, dF_{y|\underline{x}}(\theta), \tag{2.2}$$

which is assumed to exist. Then the distance measure based on the mean squared error of prediction is defined as:

$$MSE(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta)) = E^o[y - E_\theta(y|\underline{x})]^2$$

$$= E^o_{\underline{x}}\{E^o_{y|\underline{x}}[y - E_\theta(y|\underline{x})]\}^2 \tag{2.3}$$

where $E^o(\cdot)$ indicates that the expectation is evaluated with respect to the true distribution H^o of $(y,\underline{x})$. It can easily be shown that an equivalent form for (2.3) is:

$$MSE(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta)) = E^o_{\underline{x}}[var^o(y|\underline{x})] + E^o_{\underline{x}}[E^o(y|\underline{x}) - E_\theta(y|\underline{x})]^2 \tag{2.4}$$

where the first term is independent of θ, and $E^o(y|\underline{x})$ and $var^o(y|\underline{x})$ denote the true conditional expectation and variance of y given $\underline{x}$.

Then the distance between the true conditional distribution $H^o_{y|\underline{x}}$ and a specified conditional model $F_\theta = \{F_{y|\underline{x}}(\theta); \theta \in \Theta\}$ is defined by:

$$MSE(H^o_{y|\underline{x}}, F_\theta) \equiv \inf_{\theta\in\Theta} MSE(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta))$$

$$= \inf_{\theta\in\Theta} E^o[y - E_\theta(y|\underline{x})]^2. \tag{2.5}$$

Given appropriate regularity conditions (see, e.g., White (1981)), there will exists a θ^+ in Θ that minimizes $MSE(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta))$.[2] In such a case, the adequacy of the model F_θ is evaluated by:

$$MSE(H^o_{y|\underline{x}}, F_\theta) = E^o[y - E_{\theta^+}(y|\underline{x})]^2, \tag{2.6}$$

where

$$\theta^+ = \underset{\theta\in\Theta}{\operatorname{argmin}}\, E^o[y - E_\theta(y|\underline{x})]^2. \tag{2.7}$$

The second measure of adequacy of a conditional model is based on the Kullback-Leibler (1951) Information Criterion (KLIC). Specifically, the distance between the true conditional distribution $H^o_{y|\underline{x}}$ and a given conditional distribution $F_{y|\underline{x}}(\theta)$ is defined as:

$$KLIC(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta)) = E^o\left[\log \frac{h^o(y|\underline{x})}{f(y|\underline{x};\theta)}\right] \tag{2.8}$$

where $h^o(\cdot|\cdot)$ and $f(\cdot|\cdot;\theta)$ denote the conditional densities of $H^o_{y|\underline{x}}$ and $F_{y|\underline{x}}(\theta)$ with respect to a common measure ν_y. In our case, ν_y will be the Lebesgue measure since y takes its value in $\mathbb{R}$. The conditional density $f(\cdot|\cdot;\theta)$ clearly exists since we are considering normal linear regression models. On the other hand, to ensure the existence of the true density $h^o(\cdot|\cdot)$ we make the following assumption, which will be useful later on. Let $H^o_{\underline{x}}$ be the true marginal distribution of $\underline{x}$.

Assumption A2: For $H^o_{\underline{x}}$-almost all $\underline{x}$, $H^o_{y|\underline{x}}(\cdot|\underline{x})$ admits a strictly positive density $h^o(\cdot|\underline{x})$ with respect to the Lebesgue measure ν_y.

As for the previous distance measure, the distance between the true conditional distribution $H^o_{y|\underline{x}}$ and a specified conditional model F_θ is defined by:

$$KLIC(H^o_{y|\underline{x}}, F_\theta) \equiv \inf_{\theta\in\Theta} KLIC(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta))$$

$$= E^o[\log h^o(y|\underline{x})] - \inf_{\theta\in\Theta} E^o[\log f(y|\underline{x};\theta)].^3 \quad (2.9)$$

Given appropriate regularity conditions (see, e.g., White (1982a)), there exists a unique θ^* in Θ, called the pseudo-true parameters (see, e.g., Sawa (1978)), that minimizes $KLIC(H^o_{y|\underline{x}}, F_{y|\underline{x}}(\theta))$. In this case, the adequacy of a model F_θ is evaluated by:

$$KLIC(H^o_{y|\underline{x}}, F_\theta) = E^o[\log h^o(y|\underline{x})] - E^o[\log f(y|\underline{x};\theta^*)], \quad (2.10)$$

where

$$\theta^* = \underset{\theta\in\Theta}{\operatorname{argmax}}\, E^o[\log f(y|\underline{x};\theta)]. \quad (2.11)$$

Each of the above two measures of distance can naturally be used to construct a theoretical model selection criterion. Let F_θ and $G_\gamma \equiv \{G_{y|\underline{x}}(\gamma); \gamma \in \Gamma\}$ be two competing conditional models for y given $\underline{x}$. Then, using the mean squared error distance (2.6), we say that:

F_θ is MSE-better than G_γ iff $\Delta MSE(F_\theta, G_\gamma) > 0$,

F_θ is MSE-equivalent to G_γ iff $\Delta MSE(F_\theta, G_\gamma) = 0$,

F_θ is MSE-worse than G_γ iff $\Delta MSE(F_\theta, G_\gamma) < 0$,

where

$$\Delta MSE(F_\theta, G_\gamma) \equiv E^o[y - E_{\gamma^+}(y|\underline{x})]^2 - E^o[y - E_{\theta^+}(y|\underline{x})]^2 \quad (2.12)$$

where θ^+ is defined by (2.7) and γ^+ by a similar equation for the model G_γ. Using (2.4), an equivalent expression is:

$$\Delta MSE(F_\theta, G_\gamma) = E^o_{\underline{x}}[E^o(y|\underline{x}) - E_{\gamma^+}(y|\underline{x})]^2 - E^o_{\underline{x}}[E^o(y|\underline{x}) - E_{\theta^+}(y|\underline{x})]^2. \quad (2.13)$$

Equation (2.13) shows that the definitions of MSE-better, MSE-equivalent, and MSE-worse are in fact identical to those proposed by White and Olson (1979).

Alternatively, using the KLIC distance (2.10), we say that

F_θ is KLIC-better than G_γ iff $\Delta KLIC(F_\theta, G_\gamma) > 0$,

F_θ is KLIC-equivalent to G_γ iff $\Delta KLIC(F_\theta, G_\gamma) = 0$,

F_θ is KLIC-worse than G_γ iff $\Delta KLIC(F_\theta, G_\gamma) < 0$,

where

$$\Delta KLIC(F_\theta, G_\gamma) \equiv E^o[\log f(y|\underline{x};\theta^*)] - E^o[\log g(y|\underline{x},\gamma^*)] \qquad (2.14)$$

and θ^* and γ^* are the pseudo-true parameters for the conditional models F_θ and G_γ. The latter definitions were those adopted in Vuong (1985).

The essential difference between the above two sets of definitions follows from the fact that the model selection criterion based on the MSE prediction takes only into account the discrepancy between the true conditional expectation $E^o(y|\underline{x})$ and the "best" conditional mean $E_{\theta^+}(y|\underline{x})$, while the model selection criterion based on the KLIC takes into account the discrepancy between the whole true conditional density $h^o(\cdot|\cdot)$ and the "best" conditional density $f(\cdot|\cdot;\theta^*)$.[4] Thus a model which is better according to the MSE criterion is not necessarily better according to the KLIC. An important exception, however, is when one conditional model is correctly specified, i.e., when one conditional model contains the true conditional distribution $H^o_{y|\underline{x}}$. Indeed from (2.13), (2.14), and

Jensen's inequality, it follows that a correctly specified model is always at least as good as any other models according to either model selection criterion. This latter property is highly desirable and justifies the use of the above model selection criteria in the search of a correctly specified model.

When the competing models are linear regression models with normal errors, the definitions based on the MSE of prediction and on the KLIC are, however, identical as we shall see below. Let $Var^o(y,\underline{x})$ denote the true covariance matrix of y and $\underline{x}$, which we partition as follows:

$$Var^o(y,\underline{x}) = \begin{bmatrix} \sigma^o_{yy} & \sum^o_{y\underline{x}} \\ \sum^o_{\underline{x}y} & \sum^o_{\underline{x}\underline{x}} \end{bmatrix} \qquad (2.15)$$

The next assumption rules out perfect multicollinearity among all the exogenous variables $\underline{x}$.

Assumption A3: $Var^o(y,\underline{x})$ is finite, and $\sum^o_{\underline{x}\underline{x}}$ is non-singular.

Assumption A3 implies that the true means μ^o_y and $\mu^o_{\underline{x}}$ of y and $\underline{x}$ exist. The next lemma relates the values of θ^+ defined in (2.7) to the pseudo-true values θ^* defined in (2.11) when the model is a normal linear regression model with explanatory variables $\underline{x}_S \subseteq \underline{x}$. This follows by noticing that for such a model we have:

$$E^o[\log d(y|\underline{x}_S;\theta_S)] = \frac{-1}{2\sigma^2_S} E^o[y - \lambda_c - \underline{x}_S'\underline{\lambda}_S]^2 - \frac{1}{2}\log \sigma^2_S - \frac{1}{2}\log 2\pi. \qquad (2.16)$$

Lemma 2.1: Let M_s be a normal linear regression model for y give $\underline{x}$ with explanatory variables $\underline{x}_s$. Then, given A2 - A3,

$$\lambda_c^* = \lambda_c^+ = \mu_y^o - \sum_{y\underline{x}_s}^o (\sum_{\underline{x}_s\underline{x}_s}^o)^{-1}\mu_{\underline{x}_s}^o \tag{2.17}$$

$$\underline{\lambda}_s^* = \underline{\lambda}_s^+ = (\sum_{\underline{x}_s\underline{x}_s}^o)^{-1}\sum_{\underline{x}_s y}^o \tag{2.18}$$

$$\sigma_s^{*2} = \sigma_{yy}^o - \sum_{y\underline{x}_s}^o (\sum_{\underline{x}_s\underline{x}_s}^o)^{-1}\sum_{\underline{x}_s y}^o \tag{2.19}$$

where $\mu_{\underline{x}_s}^o$, $\sum_{\underline{x}_s y}^o$, and $\sum_{\underline{x}_s\underline{x}_s}^o$ are the true means and covariances corresponding to the explanatory variables $\underline{x}_s$.[5]

The next corollary gives a simple interpretation of the pseudo-true values θ^* under additional assumptions on $H_{y|\underline{x}}^o$. It is known and stated here for further reference.

Corollary 2.2: In addition to the assumptions of Lemma 2.1, suppose that the true conditional mean $E^o(y|\underline{x})$ is linear in $\underline{x}$ and the true conditional variance $var^o(y|\underline{x})$ is independent of $\underline{x}$, then

$$E^o(y|\underline{x}) = \lambda_c^* + \underline{x}_s'\underline{\lambda}_s^*, \tag{2.20}$$

$$Var^o(y|\underline{x}) = \sigma_s^{*2}. \tag{2.21}$$

We are now in a position to establish the equivalence between the MSE criterion and the KLIC for linear regression models. Let $\underline{x}_f$ and $\underline{x}_g$ be two subsets (not necessarily disjoints) of $\underline{x}$. We consider discriminating between two normal linear regression models:

$$M_f \equiv \{N(\lambda_c^f + \underline{x}_f'\underline{\lambda}_f, \sigma_f^2); \theta_f \equiv (\lambda_c^f, \underline{\lambda}_f', \sigma_f^2)' \in \mathbb{R}^{\ell_f+1} \times \mathbb{R}_+\}, \tag{2.22}$$

$$M_g \equiv \{N(\lambda_c^g + \underline{x}_g'\underline{\lambda}_g, \sigma_g^2); \theta_g \equiv (\lambda_c^g, \underline{\lambda}_g', \sigma_g^2)' \in \mathbb{R}^{\ell_g+1} \times \mathbb{R}_+\}. \tag{2.23}$$

From Lemma 2.1, we have:

$$\sigma_f^{*2} = \sigma_{yy}^o - \sum_{y\underline{x}_f}^o (\sum_{\underline{x}_f\underline{x}_f}^o)^{-1}\sum_{\underline{x}_f y}^o, \tag{2.24}$$

$$\sigma_g^{*2} = \sigma_{yy}^o - \sum_{y\underline{x}_g}^o (\sum_{\underline{x}_g\underline{x}_g}^o)^{-1}\sum_{\underline{x}_g y}^o. \tag{2.25}$$

Proposition 2.3: Given A2 - A3,

(i) $$\Delta MSE(M_f, M_g) = \sigma_g^{*2} - \sigma_f^{*2}, \tag{2.26}$$

(ii) $$\Delta KLIC(M_f, M_g) = \frac{1}{2}\log(\sigma_g^{*2}/\sigma_f^{*2}). \tag{2.27}$$

Proposition 2.3 shows that, when comparing normal linear regression models, the definitions of better than, equivalent to, and worse than are identical for the MSE criterion and the KLIC.

3. A SURVEY OF SOME MODEL SELECTION PROCEDURES

The quantities $MSE(H^o(y|\underset{\sim}{x}),F_{y|\underset{\sim}{x}}(\theta^+))$ and $E^o[\log f(y|\underset{\sim}{x};\theta^*)]$, which define the previous two theoretical model selection criteria, are unfortunately unknown. These quantities, which can be viewed as theoretical losses, can nonetheless be consistently estimated. In this section, we shall show that most of the current model selection criteria can be thought of as estimates of either theoretical loss. Our treatment differs from the usual one given in standard textbooks (see, e.g., Chow (1983), Judge et al. (1985)) which introduce these model selection criteria as estimates of the risk $E^o_{\hat{\theta}}[MSE(H^o(y|\underset{\sim}{x}),F_{y|\underset{\sim}{x}}(\hat{\theta}))]$ or $E^o_{\hat{\theta}}\,E^o[\log f(y|\underset{\sim}{x};\hat{\theta})]$ associated with an estimator $\hat{\theta}$ of θ^+ or θ^*. Thus our rather classical treatment of model selection will not allow us to discuss the alternative Bayesian solution based on Jeffrey's posterior probability criterion which has been especially studied by Zellner (1971) and Leamer (1979) in the linear regression context.

To simplify the notation and the following algebra, we assume from now on that:

Assumption A4: $\mu^o_y = 0$, $\mu^o_{\underset{\sim}{x}} = 0$.

Then, the competing normal linear regression models for y given $\underset{\sim}{x}$ that we consider are of the form:

$$M_s = \{N(\underset{\sim}{x}_s'\underset{\sim}{\lambda}_s,\sigma_s^2);\theta_s \equiv (\underset{\sim}{\lambda}_s',\sigma_s^2)' \in \mathbb{R}^{\ell_s} \times \mathbb{R}_+\}, \tag{3.1}$$

where $\underset{\sim}{x}_s$ is a subset of $\underset{\sim}{x}$, and s = f,g. That is, we exclude the constant term.

Given a random sample of size n (Assumption A1), the log-likelihood function for the model M_s is

$$L_n^s(\theta_s) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_s^2 - \frac{1}{2\sigma_s^2}(Y - X_s\underset{\sim}{\lambda}_s)'(Y - X_s\underset{\sim}{\lambda}_s) \tag{3.2}$$

where we use the convention that a capital letter denotes a matrix of the n observations on the corresponding random variables. It is well-known that the maximum-likelihood (ML) estimator of θ_s is given by:

$$\hat{\underset{\sim}{\lambda}}_s = (X_s'X_s)^{-1}X_s'Y \tag{3.3}$$

$$\hat{\sigma}_s^2 = (1/n)\sum_{t=1}^{n}\hat{e}_{st}^2 \tag{3.4}$$

where

$$\hat{e}_{st} \equiv y_t - \underset{\sim}{x}_{st}'\hat{\underset{\sim}{\lambda}}_s. \tag{3.5}$$

Hence, for s = f,g,

$$L_n^s(\hat{\theta}_s) = -(n/2)\log\hat{\sigma}_s^2 - (n/2)\log 2\pi - n/2. \tag{3.6}$$

In addition, from the theory of quasi-ML estimation (see, e.g., White (1982a)), it is known that under A1

$$\hat{\underset{\sim}{\lambda}}_s \xrightarrow{a.s.} \underset{\sim}{\lambda}_s^*,\ \hat{\sigma}_s^2 \xrightarrow{a.s} \sigma_s^{*2},\ s = f,g, \tag{3.7}$$

whether or not the model M_s is correctly specified, i.e., whether or not $H^o_{y|\underset{\sim}{x}} \in M_s$, $s = f,g$.

We now begin with model selection criteria based on the MSE. From (2.26) and (3.7), it follows that the statistic

$$\Delta\sigma^2 \equiv \hat{\sigma}^2_g - \hat{\sigma}^2_f \tag{3.8}$$

is a consistent estimator of the MSE criterion $\Delta MSE(M_f,M_g)$ whether or not the competing models M_f and M_g are correctly specified. Such a statistic was proposed by White and Olson (1979) to discriminate between M_f and M_g by testing the null hypothesis that the models are (MSE) equivalent.[6,7]

There exist, however, many other consistent estimators of $\Delta MSE(M_f,M_g)$, of which some may have better small sample properties than the natural statistic $\Delta\sigma^2$. For instance consider the statistics

$$\Delta C_p = [\hat{\sigma}^2_g + \frac{2\ell_g}{n-\ell}\hat{\sigma}^2] - [\hat{\sigma}^2_f + \frac{2\ell_f}{n-\ell}\hat{\sigma}^2] \tag{3.9}$$

$$\Delta PC = \frac{n+\ell_g}{n-\ell_g}\hat{\sigma}^2_g - \frac{n+\ell_f}{n-\ell_f}\hat{\sigma}^2_f \tag{3.10}$$

where $\hat{\sigma}$ is the ML estimator of σ^2 in the comprehensive normal linear regression model for y given $\underset{\sim}{x}$:

$$M_{f\vee g} = \{N(\underset{\sim}{x}'\underset{\sim}{\lambda},\sigma^2), \theta \equiv (\underset{\sim}{\lambda}',\sigma^2)' \in \mathbb{R}^{\ell} \times \mathbb{R}_+\}. \tag{3.11}$$

It can easily be shown that ΔC_p and ΔPC correspond to Mallows (1973) C_p criterion and to Amemiya (1980) PC criterion respectively. For

fixed explanatory variables, both ΔC_p and ΔPC are known to be unbiased estimators of the difference in MSE risk $E^o_{\hat{\underset{\sim}{\lambda}}_g}[E^o(y - \underset{\sim}{x}_g'\underset{\sim}{\lambda}_g)^2] - E^o_{\hat{\underset{\sim}{\lambda}}_f}[E^o(y - \underset{\sim}{x}_f'\hat{\underset{\sim}{\lambda}}_f)^2]$, the former under the assumption that the comprehensive model $M_{f\vee g}$ is correctly specified, and the latter under the assumption that M_f and M_g are both correctly specified (see, e.g., Judge et al. (1985, Chapter 21)). In any case, it follows from (3.7), (3.9), (3.10) and the almost sure convergence of $\hat{\sigma}^2$ to $\sigma^{*2} \equiv \sigma^o_{yy} - \sum^o_{y\underset{\sim}{x}}(\sum^o_{\underset{\sim}{xx}})^{-1}\sum^o_{\underset{\sim}{x}y}$, that ΔC_p and ΔPC are consistent estimators of $\Delta MSE(M_f,M_g) = \sigma^{*2}_g - \sigma^{*2}_f$ whether or not M_f, M_g, and $M_{f\vee g}$ are correctly specified. Indeed, from (3.8) - (3.10) we have:

$$\Delta C_p = \Delta\sigma^2 + O_p(n^{-1}), \tag{3.12}$$

$$\Delta PC = \Delta\sigma^2 + O_p(n^{-1}). \tag{3.13}$$

Hence $n^{1/2}\Delta C_p$ and $n^{1/2}\Delta PC$ will have the same asymptotic distribution as $n^{1/2}\Delta\sigma^2$ whenever $n^{1/2}\Delta\sigma^2$ converges in distribution to a limit. Such an approximation is useful since the exact finite sample distributions of $\Delta\sigma^2$, ΔC_p, and ΔPC are difficult to obtain especially when $M_{f\vee g}$ is misspecified.[8]

Next, we turn to model selection criteria based on the KLIC. From (2.27) and (3.7), a natural consistent estimator of $\Delta KLIC(M_f,M_g)$ is:

$$\frac{1}{n}LR_n(\hat{\theta}_f,\hat{\theta}_g) \equiv \frac{1}{2}\log\,(\hat{\sigma}_g^2/\hat{\sigma}_f^2)$$

$$= \frac{1}{n}[L_n^f(\hat{\theta}_f) - L_n^g(\hat{\theta}_g)], \tag{3.14}$$

where the second equation follows from (3.6). In Vuong (1985), we derived the asymptotic distribution of the likelihood-ratio (LR) statistic under general conditions. This approximation was then used to construct some LR-based tests of the null hypothesis that the models M_f and M_g are KLIC-equivalent. It will be used in the next sections when the competing models are linear regression models.

As for the theoretical criterion based on the MSE, there exist other consistent estimators of $\Delta KLIC(M_f,M_g)$. In particular, we have:

$$\frac{1}{n}\Delta AIC \equiv \frac{1}{n}[L_n^f(\hat{\theta}_f) - \ell_f] - \frac{1}{n}[L_n^g(\hat{\theta}_g) - \ell_g] \tag{3.15}$$

$$\frac{1}{n}\Delta BIC \equiv \frac{1}{n}[L_n^f(\hat{\theta}_f) - \frac{\hat{\sigma}_f^2}{\hat{\sigma}_f^2}(\ell_f + 1 - \frac{\hat{\sigma}_f^2}{\hat{\sigma}_f^2})] - \frac{1}{n}[L_n^g(\hat{\theta}_g) - \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2}(\ell_g + 1 - \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2})] \tag{3.16}$$

$$\frac{1}{n}\Delta SIC \equiv \frac{1}{n}[L_n^f(\hat{\theta}_f) - \frac{1}{2}\ell_f \log n] - \frac{1}{n}[L_n^g(\hat{\theta}_g) - \frac{1}{2}\ell_g \log n]. \tag{3.17}$$

Criterion (3.15) corresponds to Akaike (1973) information criterion. Criterion (3.16) corresponds to Sawa (1978) information criterion for normal linear regression models.[9] Criterion (3.17) corresponds to Schwarz (1978) formula for discriminating between models. These criteria were derived under different assumptions. For instance, ΔAIC was derived under the assumption that both models M_f and M_g are

correctly specified, while ΔBIC was derived under the assumption that the comprehensive model $M_{f \vee g}$ is correctly specified. From (2.27), (3.6), (3.7) and (3.15) - (3.17), it is clear, however, that $n^{-1}\Delta AIC$, $n^{-1}\Delta BIC$, and $n^{-1}\Delta SIC$ are all strongly consistent estimators of $\Delta KLIC(M_f,M_g) = \frac{1}{2}\log(\sigma_g^{*2}/\sigma_f^{*2})$. In addition, we have under general conditions:

$$\frac{1}{n}\Delta AIC = \frac{1}{n}LR_n(\hat{\theta}_f,\hat{\theta}_g) + O_p(n^{-1}), \tag{3.18}$$

$$\frac{1}{n}\Delta BIC = \frac{1}{n}LR_n(\hat{\theta}_f,\hat{\theta}_g) + O_p(n^{-1}), \tag{3.19}$$

$$\frac{1}{n}\Delta SIC = \frac{1}{n}LR_n(\hat{\theta}_f,\hat{\theta}_g) + O_p(n^{\alpha-1}) \tag{3.20}$$

for any $\alpha > 0$. Hence $n^{-1/2}\Delta AIC$, $n^{-1/2}\Delta BIC$, $n^{-1/2}\Delta SIC$ will have the same asymptotic distribution as $n^{-1/2}LR_n(\hat{\theta}_f,\hat{\theta}_g)$ whenever $n^{-1/2}LR_n(\hat{\theta}_f,\hat{\theta}_g)$ converges in distribution to a limit, as it will be the case in the next sections.

The previous discussion suggests that a classical approach for choosing between two competing models M_f and M_g is to test the null hypothesis

$$H_o: \quad M_f \text{ is equivalent to } M_g \tag{3.21}$$

against either one of the alternatives

$$H_f: \quad M_f \text{ is better than } M_g, \tag{3.22}$$

$$H_g: \quad M_g \text{ is better than } M_f. \tag{3.23}$$

As shown in Section 2, the MSE criterion and the KLIC are equivalent when comparing normal linear regression models. Thus one can equivalently test H_o: $\Delta MSE(M_f,M_g) = 0$ against H_f: $\Delta MSE(M_f,M_g) > 0$ [or H_g: $\Delta MSE(M_f,M_g) < 0$], or H_o: $\Delta KLIC(M_f,M_g) = 0$ against H_f: $\Delta KLIC(M_f,M_g) > 0$ [or H_g: $\Delta KLIC(M_f,M_g) < 0$]. Following Vuong (1985), we shall use the natural LR statistic $LR_n(\hat{\theta}_f,\hat{\theta}_g)$, but clearly any one of the previous statistics $\Delta\sigma^2$, ΔC_p, ΔPC, ΔAIC, ΔBIC, and ΔSIC can be used instead because of their small sample properties.

We shall mainly study the case where the linear regression models M_f and M_g are non-nested, and we shall propose some model selection tests under different information structures. Namely, we shall successively treat (i) the general case where none of the competing models is correctly specified, (ii) the case where the comprehensive model $M_{f \vee g}$ is correctly specified, and (iii) the case where at least one model is correctly specified. The classical case where the linear regression models are nested will be discussed only briefly in the conclusion.

4. THE GENERAL CASE

Let M_f and M_g be two normal linear regression models of y given $\underline{x}$, i.e., models of the form (3.1). Let $\underline{x}_s$ be the vector of included explanatory variables in the model M_s, s = f,g. The models M_f and M_g are assumed to be non-nested. Thus, we assume:

Assumption A5: $\underline{x}_f \not\subset \underline{x}_g$ and $\underline{x}_g \not\subset \underline{x}_f$.

It is convenient to partition $\underline{x}_f$ and $\underline{x}_g$ into some common explanatory variables and some variables specific to M_f and M_g:

$$\underline{x}_f' = (x',z'), \quad \underline{x}_g' = (x',w') \tag{4.1}$$

where x, z, and w are respectively k, p, and q dimensional vectors. Thus $\ell_f = k + p$ and $\ell_g = k + q$. The coefficient vectors $\underline{\lambda}_f$ and $\underline{\lambda}_g$ are partitioned accordingly into:

$$\underline{\lambda}_f' = (\alpha',\beta'), \quad \underline{\lambda}_g' = (\gamma',\delta') \tag{4.2}$$

Then A5 is equivalent to the assumption that $p \neq 0$ and $q \neq 0$. Without loss of generality, we shall assume throughout that $p \geq q$ and that the union of $\underline{x}_f$ and $\underline{x}_g$ is equal to $\underline{x}$ so that $k + p + q = \ell$.

Strictly speaking, linear regression models can never be strictly non-nested since they must have some common conditional distributions for y give $\underline{x}$ (see Vuong (1985, Definition 5.1)). Indeed, it is easy to see from (3.1) that M_f and M_g must both contain the non-empty class of conditional distributions for y give $\underline{x}$:

$$M_o = \{N(\underline{x}'\underline{\lambda};\sigma^2); \underline{\lambda} = 0, \sigma^2 \in \mathbb{R}_+\} \tag{4.3}$$

Hence, linear regression models can only be either overlapping (Vuong (1985, Definition 6.1)) or nested (Vuong (1985, Definition 7.1)).

The fact that $M_f \cap M_g \neq \emptyset$ even in the non-nested case has not often been recognized in the literature, and in fact much complicates the derivation of some classical tests of the null hypothesis that the models M_f and M_g are MSE or KLIC equivalent. Indeed, as shown in

Vuong (1985, Theorem 3.5), the asymptotic distribution of the LR statistic as well as the speed at which it converges to that distribution crucially depends on whether or not the closest distribution in M_f to the true distribution $H^o_{y|\underset{\sim}{x}}$ is H^o-almost surely identical to the closest distribution in M_g to $H^o_{y|\underset{\sim}{x}}$, i.e., on whether or not

$$H_o^\omega: \; d_f(\cdot|\cdot;\theta_f^*) = d_g(\cdot|\cdot;\theta_g^*) \; H^o\text{-almost surely,} \tag{4.4}$$

holds, where $d_s(\cdot|\cdot;\theta_s)$ denotes the univariate conditional normal density of y given $\underset{\sim}{x}_s$ with parameter θ_s. For s = f,g, let $e_s = y - \underset{\sim}{x}_s'\underset{\sim}{\lambda}_s^*$. Since $\text{var}^o(e_s) = \sigma_s^{*2}$, then it can readily be shown that the null hypothesis H_o^ω can be equivalently rewritten as:

$$H_o^\omega: \; e_f^2 = e_g^2 \quad H^o\text{-almost surely.} \tag{4.5}$$

Then, applying Vuong (1985) Theorem 5.2 to the normal linear regression models we obtain:

Proposition 4.1: Given A1 - A5,

(i) under $H_o - H_o^\omega$, $T_{fg} \xrightarrow{D} N(0,1)$,

(ii) under H_f, $T_{fg} \xrightarrow{a.s.} +\infty$,

(iii) under H_g, $T_{fg} \xrightarrow{a.s.} -\infty$,

where

$$T_{fg} = \frac{\log\left(\sum_{t=1}^n \hat{e}_{gt}^2 / \sum_{t=1}^n \hat{e}_{ft}^2\right)}{\left[\sum_{t=1}^n \left[\frac{\hat{e}_{gt}^2}{\sum_{t=1}^n \hat{e}_{gt}^2} - \frac{\hat{e}_{ft}^2}{\sum_{t=1}^n \hat{e}_{ft}^2}\right]^2\right]^{1/2}} \tag{4.6}$$

The statistic T_{fg} is nothing else than $n^{-1/2}LR_n(\hat{\theta}_f,\hat{\theta}_g)/\hat{\omega}_n$ where

$$\hat{\omega}_n^2 \equiv \frac{n}{4}\sum_{t=1}^n \left[\frac{\hat{e}_{gt}^2}{\sum_{t=1}^n \hat{e}_{gt}^2} - \frac{\hat{e}_{ft}^2}{\sum_{t=1}^n \hat{e}_{ft}^2}\right]^2. \tag{4.7}$$

It is important to note that we necessarily have $H_o^\omega \subset H_o$. Thus assuming that H_o^ω does <u>not</u> hold, Proposition 4.1 - (i) gives us a simple asymptotically normal test of the null hypothesis that the normal linear regression models M_f and M_g are equivalent. The test is <u>directional</u>, and Parts (ii) and (iii) ensure that the test is consistent against the alternatives H_f and H_g.

As mentioned in Vuong (1985, Section 5), the statistic T_{fg} can in general be obtained from an additional linear regression. Let

$$m_t = \frac{\hat{e}_{gt}^2}{\sum_{t=1}^n \hat{e}_{gt}^2} - \frac{\hat{e}_{ft}^2}{\sum_{t=1}^n \hat{e}_{ft}^2} + \log\frac{\sum_{t=1}^n \hat{e}_{gt}^2}{\sum_{t=1}^n \hat{e}_{ft}^2}.$$

Then, it can easily be shown that T_{fg} is equal to $[n(n-1)]^{-1/2}$ times the usual t-statistic on the constant term in a linear regression of m_t on only the constant term.

However, the previous statistic can only be used to test $H_o - H_o^\omega$. Since H_o^ω is part of the null hypothesis H_o, it is also necessary to test H_o^ω in order to determine whether M_f and M_g are equivalent. In Vuong (1985), we propose the following two-step procedure: Test H_o^ω against its alternative H_A^ω: if H_o^ω cannot be rejected then the two models are equivalent, otherwise test $H_o - H_o^\omega$ using the statistic T_{fg} as indicated in the previous paragraph. As shown there, if α_1 and α_2 are the asymptotic significance level of these tests, then the asymptotic significance level α of this sequential procedure as a test of the null hypothesis of interest H_o is <u>not larger</u> than $\max(\alpha_1,\alpha_2)$. Hence, if $\alpha_1 = \alpha_2 = 10\%$, then $\alpha \leq 10\%$.

The remainder of this section considers various ways for testing H_o^ω. Let

$$\omega^2 = \mathrm{Var}^o\left[\log \frac{f_f(y|\underset{\sim}{x}_f;\theta_f^*)}{f_g(y|\underset{\sim}{x}_g;\theta_g^*)}\right],$$

$$= \frac{1}{4}E^o\left[\frac{e_f^2}{\sigma_f^{*2}} - \frac{e_g^2}{\sigma_g^{*2}}\right]^2. \qquad (4.8)$$

Recall that $\underset{\sim}{\lambda}_f^* \equiv (\alpha^{*\prime},\beta^{*\prime})'$ and $\underset{\sim}{\lambda}_g^* \equiv (\gamma^{*\prime},\delta^{*\prime})'$ are the pseudo-true parameters for $\underset{\sim}{\lambda}_f$ and $\underset{\sim}{\lambda}_g$ in the models M_f and M_g respectively. We shall also consider the comprehensive normal linear regression model

for y given $\underset{\sim}{x} = (x',z',w')'$ defined in (3.11). Let $\underset{\sim}{\lambda}^* = (\lambda_x^{*\prime},\lambda_z^{*\prime},\lambda_w^{*\prime})'$ be the pseudo-true parameters corresponding to this model. (These can be obtained from Lemma 2.1 by setting $\underset{\sim}{x}_s = \underset{\sim}{x}$.) Then, let $e_{f\vee g} = y - \underset{\sim}{x}'\underset{\sim}{\lambda}^*$.

<u>Lemma 4.2</u>: Given A2 - A5, the null hypothesis H_o^ω is equivalent to either <u>one</u> of the following statements:

(i) $\omega^2 = 0$,

(ii) $\beta^* = 0$ and $\delta^* = 0$,

(iii) $\lambda_z^* = 0$ and $\lambda_w^* = 0$.

That H_o^ω is equivalent to $\omega^2 = 0$ is a general result (see Vuong (1985, Lemma 4.1)). It is easy to see that each of the latter two statements implies H_o^ω. These implications do not depend on Assumption A2. On the other hand, as the example in the Appendix shows, their converses crucially depend on A2 which states that the true conditional distribution $H^o_{y|\underset{\sim}{x}}$ has a strictly positive density with respect to the Lebesgue measure ν_y. In particular, y cannot be a discrete variable or have mass points.

Lemma 4.2 is, however, intuitively desirable. Indeed, rewriting (4.4) in the linear regression form, Part (ii) says that the conditional distributions for y given $\underset{\sim}{x} = (x',z',w')'$ defined by:

$$y = x'\alpha^* + z'\beta^* + e_f, \quad e_f \sim N(0,\sigma_f^{*2}) \qquad (4.9)$$

$$y = x'\gamma^* + w'\delta^* + e_g, \quad e_g \sim N(0,\sigma_g^{*2}) \qquad (4.10)$$

are H^o-almost surely identical if and only if $\beta^* = \delta^* = 0$, or equivalently if and only if $\lambda_z^* = \lambda_w^* = 0$ in the comprehensive conditional distribution for y given $(x',z',w')'$:

$$y = x'\lambda_x^* + z'\lambda_z^* + w'\lambda_w^* + e_{f\vee g},\ e_{f\vee g} \sim N(0,\sigma^{*2}_{f\vee g}). \tag{4.11}$$

If either one of this conditional holds, then

$$\alpha^* = \gamma^* = \lambda_x^* \tag{4.12}$$

$$\sigma_f^{*2} = \sigma_g^{*2} = \sigma^{*2}_{f\vee g} \tag{4.13}$$

$$e_f = e_g = e_{f\vee g} \tag{4.14}$$

(see the proof of Lemma 4.2).[10]

Lemma 4.2 allows us to test H_o^ω in various ways. For instance, using Part (i), we can test H_o^ω by using the estimator $\hat{\omega}_n^2$ defined in (4.7). This is the general procedure proposed in Vuong (1985, Theorem 4.4) where it is shown that the statistic $n\hat{\omega}_n^2$ is asymptotically distributed under H_o^ω as a weighted sum of chi-squares with weights equal to the squares of the eigenvalues of the matrix

$$W = \begin{bmatrix} -B_f A_f^{-1} & ; & -B_{fg} A_g^{-1} \\ B_{gf} A_f^{-1} & ; & B_g A_g^{-1} \end{bmatrix} \tag{4.15}$$

where, as usual,

$$A_s \equiv E^o\left[\frac{\partial^2 \log f_s(y|\underline{x}_s;\theta_s^*)}{\partial\theta_s \partial\theta_s'}\right] \tag{4.16}$$

$$B_s \equiv E^o\left[\frac{\partial \log f_s(y|\underline{x}_s;\theta_s^*)}{\partial\theta_s} \cdot \frac{\partial \log f_s(y|\underline{x}_s;\theta_s^*)}{\partial\theta_s'}\right] \tag{4.17}$$

$$B_{fg} = B_{gf}' \equiv E^o\left[\frac{\partial \log f_f(y|\underline{x}_f;\theta_f^*)}{\partial\theta_f} \cdot \frac{\partial \log f_g(y|\underline{x}_g;\theta_g^*)}{\partial\theta_g'}\right] \tag{4.18}$$

with $\theta_s = (\underline{\lambda}_s', \sigma_s^2)'$ and s = f,g.[11] The next lemma determines these matrices under the hypothesis H_o^ω. Given Lemma 4.2, we can define:

$$e \equiv e_f = e_g = e_{f\vee g} \tag{4.19}$$

$$\sigma^{*2} = \sigma_f^{*2} = \sigma_g^{*2} = \sigma^{*2}_{f\vee g}. \tag{4.20}$$

Lemma 4.3: Given A2 - A5, under H_o^ω

$$A_s = -\frac{1}{\sigma^{*2}}\begin{bmatrix} \sum^o_{\underline{x}_s\underline{x}_s} & ; & 0 \\ 0 & ; & 1/(2\sigma^{*2}) \end{bmatrix} \tag{4.21}$$

$$B_s = \frac{1}{\sigma^{*4}}\begin{bmatrix} \mathrm{Var}^o(e\underline{x}_s) & ; & \frac{1}{2\sigma^{*2}}\mathrm{Cov}^o(e^2, e\underline{x}_s) \\ \frac{1}{2\sigma^{*2}}\mathrm{Cov}^o(e^2, e\underline{x}_s') & ; & \frac{1}{4\sigma^{*4}}\mathrm{Var}^o(e^2) \end{bmatrix} \tag{4.22}$$

$$B_{fg} = B_{gf}' = \frac{1}{\sigma^{*4}}\begin{bmatrix} \mathrm{Cov}^o(e\underline{x}_f, e\underline{x}_g') & ; & \frac{1}{2\sigma^{*2}}\mathrm{Cov}^o(e\underline{x}_f, e^2) \\ \frac{1}{2\sigma^{*2}}\mathrm{Cov}^o(e^2, e\underline{x}_g') & ; & \frac{1}{4\sigma^{*4}}\mathrm{Var}^o(e^2) \end{bmatrix} \tag{4.23}$$

In the general case where the models M_f and M_g are not necessarily correctly specified, the matrix W does not simplify since the information matrix equivalence $A_s + B_s = 0$ does not necessarily hold (see, e.g., White (1982)).[12] Determination of the eigenvalues of W is, however, important for the subsequent tests. Since W is of dimension $\ell_f + \ell_g + 2 = 2k + p + q + 2$, the matrix W has $2k + p + q + 2$ eigenvalues, which are all real (see Vuong (1985)). The next lemma is quite useful since it states that at least $2k + 2$ eigenvalues are zero, and shows how the remaining eigenvalues can be obtained. We need some additional notation. Define

$$R' = \begin{bmatrix} -\sum^o_{zx}\sum^{o-1}_{xx} & ; & I_p & ; & 0 \\ -\sum^o_{wx}\sum^{o-1}_{xx} & ; & 0 & ; & I_q \end{bmatrix} \tag{4.24}$$

$$C = \begin{bmatrix} C_{xx} & C_{xz} & C_{xw} \\ C_{zx} & C_{zz} & C_{zw} \\ C_{wx} & C_{wz} & C_{ww} \end{bmatrix}, \tag{4.25}$$

$$Q = \begin{bmatrix} Q_{zz} & Q_{zw} \\ Q_{wz} & Q_{ww} \end{bmatrix}. \tag{4.26}$$

where

$$C_{ij} = \mathrm{Cov}^o(ie, je), \tag{4.27}$$

$$Q_{ij} = \sum^o_{ij} - \sum^o_{ix}\left(\sum^o_{xx}\right)^{-1}\sum^o_{xj}, \tag{4.28}$$

for i,j = x,z,w. Define also

$$\mathrm{Diag}(Q_{zz}, -Q_{ww}) = \begin{bmatrix} Q_{zz} & 0 \\ 0 & -Q_{ww} \end{bmatrix}. \tag{4.29}$$

It can easily be shown that under A2, Q and hence $\mathrm{Diag}(Q_{zz}, -Q_{ww})$ are both non-singular.

Lemma 4.4: Given Assumptions A2 - A5, under H^ω_o, the matrix W has at least $2k + 2$ zero eigenvalues, the other p+q eigenvalues λ_ω are real and are solutions of

$$\det[R'CR - \lambda_\omega \sigma^{*2}\mathrm{Diag}(Q_{zz}, -Q_{ww})] = 0.^{13} \tag{4.30}$$

With all the eigenvalues of W being characterized, the test of H^ω_o against H^ω_A (that is, the variance test in terms of Vuong (1985)) using $n\hat{\omega}^2_n$ as the statistic has the following property:

Proposition 4.5 (Variance Test): Given Assumptions A1 - A5,

(i) under H^ω_o, $n\hat{\omega}^2_n \xrightarrow{D} M_{p+q}(\cdot;\lambda^2_\omega)$

(ii) under H^ω_A, $n\hat{\omega}^2_n \xrightarrow{a.s.} +\infty$

where $M_{p+q}(\cdot;\lambda^2_\omega)$ is a weighted sum of chi-squares with weights equal to λ^2_ω.

In practice, the eigenvalues λ_ω are unknown and must be consistently estimated. This can clearly be done by consistently estimating the unknown matrices in (4.30) by their sample analogs:

$$\hat{\Sigma}_{ij} = \frac{1}{n}\sum_{t=1}^{n} i_t j_t' \tag{4.31}$$

$$\hat{Q}_{ij} = \hat{\Sigma}_{ij} - \hat{\Sigma}_{ix}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xj} \tag{4.32}$$

$$\hat{C}_{ij} = \frac{1}{n}\sum_{t=1}^{n} \hat{e}_t^2 i_t j_t' \tag{4.33}$$

where $\hat{e}_t$ can be taken to be $y_t - x_t'\hat{\alpha}$, $y_t - x_t'\hat{\gamma}$, $y_t - x_t'\hat{\lambda}_x$, or more directly, $y_t - x_t'\hat{\alpha} - z_t'\hat{\beta}$, $y_t - x_t'\hat{\gamma} - z_t'\hat{\delta}$, $y_t - x_t'\hat{\lambda}_x - z_t'\hat{\lambda}_z - w_t'\hat{\lambda}_w$ because of Lemma 4.2 and (4.19).

Estimation of the non-zero eigenvalues λ_ω can be avoided if one knows a priori that all the non-zero eigenvalues are equal to one. As seen in the next section, this will not be the case in general. Hence the variance statistic is not necessarily chi-square distributed under H_o^ω. On the other hand the test is consistent against <u>all</u> alternatives.

As indicated in Lemma 4.2, H_o^ω is also equivalent to $\beta^* = \delta^* = 0$. Thus, a Wald test based on an appropriate quadratic form in $(\hat{\beta}', \hat{\delta}')'$ may replace the variance test. The next lemma gives the asymptotic covariance matrix of $n^{1/2}(\hat{\beta}', \hat{\delta}')'$ under the null hypothesis H_o^ω. Let $\hat{\Sigma}_{\underline{x}y} = (\hat{\Sigma}_{xy}', \hat{\Sigma}_{zy}', \hat{\Sigma}_{wy}')'$ where a quantity with a hat denotes

the sample analog of that quantity. For instance $\hat{\Sigma}_{xy} = \frac{1}{n}\sum_{t=1}^{n} x_t y_t$.

<u>Lemma 4.6</u>: Given Assumptions A1 - A5,

$$\text{(i)} \quad (\hat{\beta}', \hat{\delta}')' = \text{Diag}(\hat{Q}_{zz}^{-1}, \hat{Q}_{ww}^{-1})\hat{R}'\hat{\Sigma}_{\underline{x}y} \tag{4.34}$$

(ii) under H_o^ω, $n^{1/2}(\hat{\beta}', \hat{\delta}')' \xrightarrow{D} N(0,V)$ where

$$V = \text{Diag}(Q_{zz}^{-1}, Q_{ww}^{-1})R'CR\,\text{Diag}(Q_{zz}^{-1}, Q_{ww}^{-1}). \tag{4.35}$$

Since R has full-column rank (see Equation (4.24)), it follows that if C is nonsingular, then the asymptotic covariance matrix V is non-singular. In general, however, V will be singular. Thus , using generalized (g-) inverses, we define a Wald statistic as:

$$W_n^1 = n(\hat{\beta}, \hat{\delta}')G_n(\hat{\beta}, \hat{\delta}')' \tag{4.36}$$

where $G_n \xrightarrow{a.s.} G$ and G is a g-inverse of V, i.e., $G = V^-$ (see Moore (1977), Vuong (1986)).[14] Let $r \equiv \text{rank } V \leq p + q$ and let

$$\bar{H}_A^\omega = \{(\beta^{*\prime}, \delta^{*\prime})' \in M(V) - \{0\}\} \tag{4.37}$$

where M(V) denotes the r-dimensional manifold generated by the columns of V. Note that $\bar{H}_A^\omega \subset H_A^\omega$ since H_A^ω is equivalent to $(\beta^{*\prime}, \delta^{*\prime})' \neq 0$ (see Lemma 4.2). We have:

<u>Proposition 4.7</u>: Given Assumptions A1 - A5,

(i) under H_o^ω, for any choice of g-inverse G and for any consistent sequence G_n, $W_n^1 \xrightarrow{D} \chi^2_r$,

(ii) under $\bar{H}_A^\omega$, $W_n^1 \xrightarrow{a.s.} +\infty$.

Contrary to the variance statistic, the Wald statistic W_n^1 is always chi-square distributed under H_o^ω.[15] Hence the test based on W_n^1 is easier to carry out. On the other hand, if V is singular so that $r < p + q$, the Wald test would not be consistent against all alternatives in H_A^ω. Indeed, there would exist a $p + q - r$ dimensional manifold of alternatives against which the Wald test will not have any asymptotic power (see Vuong (1986)).

The previous Wald test requires two OLS regressions to obtain $\hat{\beta}$ and $\hat{\delta}$. Using Lemma 4.2 - (iii), we can think of testing H_o^ω by testing instead $\lambda_z^* = \lambda_w^* = 0$ in the comprehensive normal linear regression model $M_{f \vee g}$, and hence to do only one OLS regression. Let $\hat{\lambda}_z$ and $\hat{\lambda}_w$ be the OLS estimates of λ_z^* and λ_w^* for this comprehensive model.

Lemma 4.8: Given Assumptions A1 - A5,

$$\text{(i)} \quad (\hat{\lambda}_z', \hat{\lambda}_w')' = \hat{Q}^{-1}\hat{R}'\hat{\sum}_{xy} \tag{4.38}$$

(ii) under H_o^ω, $n^{1/2}(\hat{\lambda}_z', \hat{\lambda}_w')' \xrightarrow{D} N(0,W)$ where

$$W = Q^{-1}R'CRQ^{-1}. \tag{4.39}$$

Note that rank W = rank V = r. As before, we define a Wald statistic as:

$$W_n^2 = n(\hat{\lambda}_z', \hat{\lambda}_w')H_n(\hat{\lambda}_z', \hat{\lambda}_w')' \tag{4.40}$$

where $H_n \xrightarrow{a.s.} H$ and H is a g-inverse of W. The next result relates W_n^2 to W_n^1 and gives the asymptotic properties of the Wald test based on W_n^2. Let

$$\tilde{H}_A^\omega = \{(\beta^{*\prime}, \delta^{*\prime})' \in M(W) - \{0\}\} \tag{4.41}$$

Proposition 4.9: Given Assumptions A1 - A5,

(i) under H_o^ω, for any choice of G, H, and consistent sequences G_n and H_n, $W_n^1 - W_n^2 = o_p(1)$,

(ii) under H_o^ω, for any choice of H and consistent sequence H_n, $W_n^2 \xrightarrow{D} \chi^2_r$,

(iii) under $\tilde{H}_A^\omega$, $W_n^2 \to +\infty$.

Thus the second Wald test has the same asymptotic properties as the first Wald test.[16] It is asymptotically chi-square distributed under H_o^ω, but it is not consistent against all alternatives in H_A^ω if $r < p + q$.

5. THE COMPREHENSIVE MODEL IS CORRECTLY SPECIFIED

A well-known and important method for discriminating between two competing non-nested models is to construct a so-called comprehensive model that contains both competing models. This

approach was initially suggested by Cox (1961, 1962) and subsequently studied by Atkinson (1970). When the exponential combination of the competing densities is used, we obtain for the normal linear regression models M_f and M_g:

$$y = x'[(1 - \lambda)\frac{\sigma^2}{\sigma_f^2}\alpha + \lambda\frac{\sigma^2}{\sigma_g^2}\gamma] + (1 - \lambda)\frac{\sigma^2}{\sigma_f^2}z'\beta + \lambda\frac{\sigma^2}{\sigma_g^2}w'\delta + u \qquad (5.1)$$

where $u \sim N(0,\sigma^2)$, and $\sigma^{-2} \equiv (1 - \lambda)\sigma_f^{-2} + \lambda\sigma_g^{-2}$ (see, e.g., Pesaran (1982)). It can readily be shown that the model (5.1) for y given $\underline{x}$ is <u>identical</u> to the model $M_{f \vee g}$ defined in Equation (3.11).

When $\lambda = 0$, the model (5.1) reduces to M_f, and when $\lambda = 1$, it reduces to M_g. Then, assuming that the comprehensive model (5.1) [or $M_{f \vee g}$] is correctly specified, one successively tests the hypotheses $\lambda = 0$ and $\lambda = 1$ to determine which of the two models M_f and M_g is best. The comprehensive approach suffers, however, from (i) the arbitrariness in the choice of a comprehensive model, (ii) the necessity of carrying out two successive tests, (iii) the fact that all the parameters α, β, γ, δ, σ_f^2, σ_g^2, and λ are not identified, and (iv) that under $\lambda = 0$ (or $\lambda = 1$) the parameters δ (or β) do not enter into the combined density so that the LR test or LM test of $\lambda = 0$ (or $\lambda = 1$) is not applicable (see, e.g., Pesaran (1982)).

In this section, we shall retain the assumption that the comprehensive model $M_{f \vee g}$ is correctly specified, and we shall simplify the general model selection procedure of the previous section given this additional assumption. It is clear that we can still use the simple directional normal statistic T_{fg} discussed in Proposition

4.1 in order to test part of the null hypothesis H_o, namely $H_o - H_o^{\omega}$, against H_f or H_g. To test the remainder of the null hypothesis H_o, namely H_o^{ω}, we can consider the variance and the Wald tests discussed earlier in Propositions 4.5, 4.7, and 4.9. The purpose of this section is to simplify these latter tests when the comprehensive model $M_{f \vee g}$ is assumed to be correctly specified.

As a matter of fact, we shall only assume that the true conditional mean of y give $\underline{x}$ is linear in $\underline{x}$ and that the conditional variance is independent of $\underline{x}$. Formally we assume

<u>Assumption A6</u>: (a) $E^o(y|\underline{x}) = \underline{x}'\underline{\lambda}^o = x'\lambda_x^o + z'\lambda_z^o + w'\lambda_w^o$, (b) $var^o(y|\underline{x}) = \sigma_o^2$.

It is clear that A6 is weaker than the assumption that the comprehensive model $M_{f \vee g}$ is correctly specified. Therefore our results naturally apply to the latter case. The following lemma presents the implications of H_o^{ω} on the true conditional mean and variance when A6 is satisfied.

<u>Lemma 5.1</u>: Given Assumptions A2 - A6, under H_o^{ω}, we have

(i) $\lambda_z^o = \lambda_z^* = \beta^* = 0$; $\lambda_w^o = \lambda_w^* = \delta^* = 0$,

(ii) $\lambda_x^o = \lambda_x^* = \alpha^* = \gamma^*$

(iii) $\sigma_o^2 = \sigma_{f \vee g}^{*2} = \sigma_f^{*2} = \sigma_g^{*2}$.

That is, the imposition of Assumption A6 ensures that the true conditional expectation of y only depends on x under H_o^ω, consequently all the pseudo-true parameters are equivalent to their corresponding true parameters. From Lemma 5.1, it is clear that the addition of A6 will simplify the general expressions. In particular, under H_o^ω for any i,j = x, z, w,

$$C_{ij} = E^o(ij'e^2) = E^o(ij'E^o(e^2|x,z,w)) = \sigma_o^2 \sum_{ij}, \tag{5.2}$$

where $e = y - x'\alpha^* = y - x'\gamma^* = y - x'\lambda_x^o$ so that

$$C = \sigma_o^2 \sum. \tag{5.3}$$

Hence C is non-singular under H_o^ω.

In addition the matrices Q_{zz}, Q_{ww}, and Q_{zw} have a natural interpretation. For example, $Q_{ww} = Var^o(e_1)$ where $e_1 = w - a^{*}x$ and $a^* = \sum_{wx}^o (\sum_{xx}^o)^{-1}$; $Q_{zz} = Var^o(e_2)$ where $e_2 = z - b^{*}x$ and $b^* = \sum_{zx}^o (\sum_{xx}^o)^{-1}$. That is, we artificially set up two linear regression models with w and z being the two dependent variables, x being the common independent variable. Therefore, Q_{ww} is the variance of the residual for the first model while Q_{zz} is the variance of the residual for the second model. Moreover, if we artificially set up a linear regression model with e_1 being the dependent variable, e_2 being the independent variable, then $Var^o(e_3) = Q_{ww} - Q_{zw}' Q_{zz}^{-1} Q_{zw}$ where $e_3 = e_1 - c^{*}e_2$ and $c^* = Q_{zw}' Q_{zz}^{-1}$.

Lemma 5.2: Given A2 - A6, under H_o^ω, W has exactly (2k + 2) zero eigenvalues, p-rank Q_{zw} eigenvalues equal to one, and q-rank Q_{zw} eigenvalues equal to minus one. The remaining 2 rank Q_{zw} eigenvalues λ_ω, if any, are real with $0 < |\lambda_\omega| < 1$, and solve:

$$|Q_{wz}Q_{zz}^{-1}Q_{zw} - (1 - \lambda_\omega^2)Q_{ww}| = 0. \tag{5.4}$$

Putting $\mu = (1 - \lambda_\omega^2)$, Equation (5.4) can be solved by determining the eigenvalues of $Q_{ww}^{-1/2}Q_{wz}Q_{zz}^{-1}Q_{zw}Q_{ww}^{-1/2}$ or $Q_{wz}Q_{zz}^{-1}Q_{zw}Q_{ww}^{-1}$. This latter matrix has an interesting interpretation in terms of the vectors of random variables e_1 and e_2 defined above. Indeed, we can write $Q_{wz}Q_{zz}^{-1}Q_{zw}Q_{ww}^{-1}$ as $[Var^o(e_1) - Var^o(e_2)][Var^o(e_1)]^{-1}$, which can be treated as a generalized version of R^2 for the artificial linear regression model with e_1 being the dependent variable and e_2 being the independent variable. A particular case is when p = q = 1 so that $Q_{wz}Q_{zz}^{-1}Q_{zw}Q_{ww}^{-1}$ is the usual R^2 for the artificial population regression of e_1 on e_2.

From Proposition 4.5 and Lemma 5.2, we obtain:

Proposition 5.3: Given Assumptions A1 - A6,

(i) under H_o^ω, $n\hat{\omega}_n^2 \xrightarrow{D} \chi^2_{p+q-2 \text{ rank } Q_{zw}} + \sum_{i=1}^{\text{rank } Q_{zw}} \lambda_{\omega i}^2 \chi^2_{(2)}$,

where $0 < \lambda_{\omega i} < 1$.

(ii) under H_A^ω, $n\hat{\omega}_n^2 \to +\infty$.

Corollary 5.4: Given Assumptions A1 - A6, under H_o^ω, $n\hat{\omega}_n^2$ is asymptotically chi-square distributed if and only if $Q_{zw} = 0$.

Since $Q_{zw} = \sum_{zw}^o - \sum_{zx}^o (\sum_{xx}^o)^{-1} \sum_{zw}^o$, the condition $Q_{zw} = 0$ can be interpreted as requiring that the variables z and w are conditionally orthogonal given x. This is satisfied, for instance, if z is orthogonal to (x,w) so that $\sum_{zx}^o = \sum_{zw}^o = 0$, or if w is orthogonal to (x,z) so that $\sum_{wx}^o = \sum_{wz}^o = 0$. If there are no common explanatory variables x, then $Q_{zw} = 0$ is equivalent to z and w being orthogonal, i.e., $\sum_{zw}^o = 0$.[17]

Although the variance test generally involves the distribution of a weighted sum of chi-squares, following Section 4 we may also employ Wald statistics to test H_o^ω against H_A^ω where asymptotic chi-square distributions prevail. As noticed earlier, under H_o^ω, the covariance matrix C is non-singular so that the use of g-inverses in Equations (4.36) and (4.40) become unnecessary. Hence we can define the Wald statistics based on $(\hat{\beta}', \hat{\delta}')'$ and $(\hat{\lambda}_z', \hat{\lambda}_w')'$ directly as:

$$W_n^1 = n(\hat{\beta}', \hat{\delta}') V_n^{-1} (\hat{\beta}', \hat{\delta}')', \tag{5.5}$$

$$W_n^2 = n(\hat{\lambda}_z', \hat{\lambda}_w') W_n^{-1} (\hat{\lambda}_z', \hat{\lambda}_w')', \tag{5.6}$$

where V_n and W_n are consistent estimators under H_o^ω of the non-singular asymptotic covariance matrices V and W. From (4.35) and (4.39), we can clearly choose the consistent estimates obtained by replacing in these formulae the matrices Q and R by their sample analogs $\hat{Q}$ and $\hat{R}$,

and the matrix C by

$$\hat{C} = \hat{\sigma}^2 \hat{\sum}, \tag{5.7}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} \hat{e}_t^2, \tag{5.8}$$

where $\hat{e}_t$ can be taken to be $y_t - x_t'\hat{\alpha}$, $y_t - x_t'\hat{\gamma}$, $y_t - x_t'\hat{\lambda}_x$, $y_t - x_t'\hat{\alpha}' - z_t'\hat{\beta}$, $y_t - x_t'\hat{\gamma} - w_t'\hat{\delta}$, or $y_t - x_t'\hat{\lambda}_x - z_t'\hat{\lambda}_z - w_t'\hat{\lambda}_w$. For any choice of $\hat{e}_t$, we have:

Proposition 5.5: Given Assumptions A1 - A6,

(i) $W_n^1 = W_n^2 = n\hat{\sigma}^{-2} \hat{\sum}_{yx} \hat{R}\hat{Q}^{-1}\hat{R}^{-1} \hat{\sum}_{xy}$ (5.9)

(ii) under H_o^ω, any of these statistics converges in distribution to a χ^2_{p+q}

(iii) under H_A^ω, any of these statistics converges almost surely to $+\infty$.

Thus, contrary to the variance statistic, the Wald statistics are always asymptotically chi-square distributed. Moreover, unlike the general case (see Section 4), the Wald test are consistent against all alternatives in H_A^ω. Thus the Wald tests are preferable to the variance test since they are easier to carry out. Finally, let us note that A2 and A6 are automatically satisfied if the comprehensive model $M_{f \vee g}$ is correctly specified, i.e., if we assume:

Assumption A6': $H^o_{y|\underline{x}} = N(\underline{x}'\lambda^o, \sigma_o^2)$.

In this case, if $\hat{\sigma}^2$ is defined by (5.8) using the OLS residuals for the comprehensive model $M_{f \vee g}$, i.e., $e_t = y_t - x'\hat{\lambda}_x - z'\hat{\lambda}_z - w'\hat{\lambda}_w$, then we can in fact obtain, under H_o^ω, the exact small sample distribution of the Wald statistics W_n^1 and W_n^2 conditional on $\underline{X}$. Indeed, from the theory of linear regression models, we have under H_o^ω and given $\underline{X}$:

$$\frac{n - (k + p + q)}{n(p + q)} W_n^1 = \frac{n - (k + p + q)}{n(p + q)} W_n^2 \sim F(p + q, n - (k + p + q)).$$

6. AT LEAST ONE MODEL IS CORRECTLY SPECIFIED

In this section, we assume that at least one of the two competing models is correctly specified, i.e.,

Assumption A7: $H^o_{y|\underline{x}} \in M_f \cup M_g$.

First, it is worthnoting that A7 is stronger than the assumption considered in the previous section that the comprehensive model $M_{f \vee g}$ is correctly specified. This follows from the fact that $M_f \cup M_g$ is included in $M_{f \vee g}$. Second, assuming that one knows that one model is correctly specified does not mean that one knows which one is the correct model. Indeed, if this was the case, the correct model would be at least as good as the other model (see Section 2) and the model selection problem would be trivial. Third, though A7 appears to be more rhetorical than justified in practice, such an

assumption is often considered in the model selection literature where one chooses a model in a list of competing models, one of which being correct.

As noticed in Vuong (1985, Lemma 6.3), the major difficulty arising from the discrepancy between H_o^ω and H_o disappears since $H_o^\omega = H_o$ when at least one model is correctly specified. It follows that the speed at which the LR statistic converges to a limiting distribution remains constant under the null hypothesis H_o. Specifically, Theorem 6.4 in Vuong (1985) establishes that (i) under H_o, $2LR_n(\hat{\theta}_f, \hat{\theta}_g)$ converges in distribution to a weighted sum of chi-squares with weights equal to the eigenvalues of W, (ii) under H_f, $2LR_n(\hat{\theta}_f, \hat{\theta}_g) \to +\infty$, and (iii) under H_g, $2LR_n(\hat{\theta}_f, \hat{\theta}_g) \to -\infty$. Hence, in this case, we can bypass the sequential procedure of Sections 4 and 5 which is based on the variance (or Wald) tests followed by the normal LR test, and use directly the statistic $2LR_n(\hat{\theta}_f, \hat{\theta}_g)$ to choose between M_f and M_g. When the models are normal linear regressions, we have from (3.6):

$$2LR_n(\hat{\theta}_f, \hat{\theta}_g) = -(n/2)\log(\hat{\sigma}_f^2/\hat{\sigma}_g^2). \tag{6.1}$$

In addition, a simplification in W arises since under H_o, which is equal to H_o^ω, both models must be correctly specified. Hence the usual information matrix equivalence $A_s + B_s = 0$ holds so that (4.15) becomes

$$W = \begin{bmatrix} I & ; & B_{fg}B_g \\ -B_{gf}B_f & ; & -I \end{bmatrix} \tag{6.2}$$

Moreover since $H_o = H_o^\omega$, and since A7 clearly implies A6, the eigenvalues of W under H_o can be directly obtained from Lemma 5.2. Proposition 6.1 follows.

Proposition 6.1: Given A1 - A5, and A7,

(i) under H_o,

$$2LR_n(\hat{\theta}_f,\hat{\theta}_g) \xrightarrow{D} \chi^2_{p\text{-rank } Q_{zw}} - \chi^2_{q\text{-rank } Q_{zw}} + \sum_{i=1}^{\text{rank } Q_{zw}} \lambda_{\omega_i}(\chi^2_{(1)} - \chi^2_{(1)})$$

where $0 < \lambda_{\omega_i} < 1$, and λ_{ω_i} solves equation (5.4).

(ii) under H_f, $2LR_n(\hat{\theta}_f,\hat{\theta}_g) \to +\infty$,

(iii) under H_g, $2LR_n(\hat{\theta}_f,\theta_g) \to -\infty$.

As in Corollary 5.4, an interesting case is when $Q_{zw} = 0$.

Corollary 6.2: Given A1 - A5, and A7, if $Q_{zw} = 0$, then under H_o:

$$2LR_n(\hat{\theta}_f,\hat{\theta}_g) \xrightarrow{D} \chi^2_{p\text{-rank}(Q_{zw})} - \chi^2_{q\text{-rank}(Q_{zw})}. \qquad (6.3)$$

That is, $2LR_n(\hat{\theta}_f,\hat{\theta}_g)$ has an asymptotic distribution which can be decomposed as the difference between two chi-squares with degrees of freedom being p-rank(Q_{zw}) and q-rank(Q_{zw}), respectively. But, whether or not $Q_{zw} = 0$, it is worthnoting that $2LR_n(\hat{\theta}_f,\hat{\theta}_g)$ can never have an asymptotic chi-square distribution. Proposition 6.1 shows

that the proposed LR-based test is directional and consistent, and hence can be directly use to choose between M_f and M_g when at least one model is known to be correctly specified.

Finally, let us also note that, since $H_o = H_o^\omega$, one can think of testing H_o by applying the variance and Wald tests studied in the previous section. These latter tests are, however, not directional so that, in case of rejection of H_o, one cannot infer which of the two competing models is best.

7. CONCLUSION

In this paper, we propose a classical hypothesis approach for choosing between two normal linear regression models which may be both incorrectly specified. This approach is based on testing the null hypothesis H_o that the models are MSE or KLIC equivalent against the hypothesis that one model is closer to the truth. In general the procedure is sequential and consists in testing the stronger hypothesis H_o^ω that the closest distributions to the truth in the competing models are identical, and in case of rejection of H_o^ω to test the hypothesis $H_o - H_o^\omega$. The asymptotic significance level of the procedure as a test of H_o is however not larger than the maximum of the chosen asymptotic significance level of each test.

To test $H_o - H_o^\omega$, we propose a very simple directional and symmetric test based on the LR-statistic appropriately normalized which is asymptotically standard normal distributed under $H_o - H_o^\omega$. To test H_o^ω, we propose three tests based on the so-called variance

statistic and two Wald statistics that use either the coefficient estimates of both models or the coefficient estimates of a comprehensive linear regression model. The relationship and the consistency of these tests are studied. When the comprehensive linear model is correctly specified, the Wald based tests are identical and consistent against all alternatives to H_o^ω. In addition, implementation of the variance test simplifies. In particular, it becomes a chi-square test when the specific variables in the competing models are conditionally orthogonal given the common variables.

An important case where one does not need to use the above sequential procedure is when one competing model is known to be correctly specified. In this case $H_o = H_o^\omega$, and we propose a directional and symmetric test of H_o based directly on twice the LR-statistic which is distributed as a weighted sum of chi-squares under the null hypotheses H_o that the competing models are equivalent.

The previous sections were solely concerned with non-nested models. When the competing models are nested, the classical solution is to set the model selection problem within the hypothesis testing framework. Specifically, for the competing nested normal linear regression models:

$$M_f: \quad y = x'\alpha + z'\beta + e_f, \; e_f \sim N(0,\sigma_f^2),$$

$$M_g: \quad y = x'\gamma + e_g \qquad , \; e_g \sim N(0,\sigma_g^2),$$

the classical solution consists in testing H_o^θ: $\beta^* = 0$ against H_A^θ: $\beta^* \neq 0$.[18] As Vuong (1985, Lemma 7.2) showed, the classical

hypotheses H_o^θ and H_A^θ are in fact equivalent to the model selection hypotheses H_o and H_f, respectively. Thus, our model selection approach has the desirable property that it coincides with the usual classical hypothesis approach when the competing models are nested.[19] In other words, we can view our model selection approach as extending the classical nested hypothesis testing to non-nested situations.

As mentioned in the introduction, another solution which has recently attracted a lot of attention derives from Cox (1960, 1961)'s work on testing non-nested hypotheses.[20] When the competing models are nested, Cox's approach does not however coincide with the classical hypothesis approach in the sense that the implicit null hypothesis of the Cox test is not identical to the usual classical parametric null hypothesis. When the competing models are non-nested, Cox tests have been used to select among models (see, e.g., Pesaran and Deaton (1978)). The procedure is based on two successive tests designed to test the validity of each competing model. For instance, when the competing models are normal linear regression models, Pesaran (1974) showed that the numerator of the Cox-statistic for testing the validity of M_f is, in our notation, proportional to:

$$\log \frac{\hat{\sigma}_g^2}{\hat{\sigma}_f^2 + \frac{1}{n}\hat{\lambda}_f' X_f' M_g X_f \hat{\lambda}_f}$$

where $M_g = I - X_g(X_g'X_g)^{-1}X_g'$, and $Q_{zw} \neq 0$. It is well-known that there are nine possible outcomes to this procedure, three of which are asymptotically impossible (see, e.g., Dastoor (1981)). Vuong (1985)

has, however, shown that three of the remaining outcomes are indecisive in the sense that one cannot infer if one model is (strictly) better than the other. This is expected since Cox tests were initially proposed as diagnostic (or model specification) tests and not as model selection tests. On the other hand, our proposed model selection tests can be also thought as diagnostic tests. Indeed if the equivalence between M_f and M_g is rejected in favor of M_g being better, then M_f must be incorrectly specified (even though the better model M_g may still be incorrectly specified).

APPENDIX

Proof of Lemma 2.1: Since $(\lambda_c^*, \underset{\sim}{\lambda}_s^*, \sigma_s^{*2})$ and $(\lambda_c^+, \underset{\sim}{\lambda}_s^+)$ maximize (2.16) and $-E^o(y - \underset{\sim}{x}_s'\underset{\sim}{\lambda}_s - \lambda_c)^2$ respectively, the results follow immediately from the first order conditions.

Proof of Corollary 2.2: See Johnson and Kotz (1972, p.70).

Proof of Proposition 2.3: (i) Since $\lambda_c^+ = \lambda_c^*, \underset{\sim}{\lambda}_s^+ = \underset{\sim}{\lambda}_s^*$, we have $E^o(y - \underset{\sim}{x}_s'\underset{\sim}{\lambda}_s^+ - \lambda_c^+)^2 = E^o(y - \underset{\sim}{x}_s'\underset{\sim}{\lambda}_s^* - \lambda_c^*)^2 = \sigma_s^{*2}$. Therefore $\Delta MSE(M_f, M_g) = \sigma_g^{*2} - \sigma_f^{*2}$. (ii) Upon substituting $\lambda_c^*, \underset{\sim}{\lambda}_s^*$, and σ_s^{*2} into (2.10) we have:

$$E^o[\log \mathcal{d}_s(y|\underset{\sim}{x}_s;\theta_s^*)] = -\tfrac{1}{2}\log \sigma_s^{*2} - \tfrac{1}{2} - \tfrac{1}{2}\log 2\pi. \tag{A.1}$$

Consequently $\Delta KLIC(M_f, M_g) = \frac{1}{2}\log(\sigma_g^{*2}/\sigma_f^{*2})$.

Proof of Proposition 4.1: From (3.6) and (3.14), we have:

$$LR_n(\hat{\theta}_f, \hat{\theta}_g) = \sum_{t=1}^{n} \log[\mathcal{d}_f(y_t|\underset{\sim}{x}_{ft};\hat{\theta}_f)/\mathcal{d}_g(y_t|\underset{\sim}{x}_{gt};\hat{\theta}_g)]$$

$$= -\tfrac{n}{2}\log(\hat{\sigma}_g^2/\hat{\sigma}_f^2) = -\tfrac{n}{2}\log\Big(\sum_{t=1}^{n}\hat{e}_{ft}^2 \Big/ \sum_{t=1}^{n}\hat{e}_{gt}^2\Big). \tag{A.2}$$

Moreover,

$$[\log(\mathcal{d}_f(y_t|\underset{\sim}{x}_{ft};\hat{\theta}_f)/\mathcal{d}_g(y_t|\underset{\sim}{x}_{gt};\hat{\theta}_g))]^2$$

$$= [\tfrac{1}{2}\log(\hat{\sigma}_g^2/\hat{\sigma}_f^2) + \tfrac{1}{2}(\hat{e}_{gt}^2/\hat{\sigma}_g^2 - \hat{e}_{ft}^2/\hat{\sigma}_f^2)]^2$$

$$= \tfrac{1}{4}[\log(\hat{\sigma}_g^2/\hat{\sigma}_f^2)]^2 + \tfrac{1}{2}\log(\hat{\sigma}_g^2/\hat{\sigma}_f^2)(\hat{e}_{gt}^2/\hat{\sigma}_g^2 - \hat{e}_{ft}^2/\hat{\sigma}_f^2)$$

$$+ \tfrac{1}{4}(\hat{e}_{gt}^2/\hat{\sigma}_g^2 - \hat{e}_{ft}^2/\sigma_f^2)^2$$

Therefore,

$$\tfrac{1}{n}\sum_{t=1}^{n}[\log(\mathcal{J}_f(y_t|\underset{\sim}{x}_{ft};\hat{\theta}_f)/\mathcal{J}_g(y_t|\underset{\sim}{x}_{gt};\hat{\theta}_g))]^2$$

$$= \tfrac{1}{4}[\log(\hat{\sigma}_g^2/\hat{\sigma}_f^2)]^2 + \tfrac{n}{4}\sum_{t=1}^{n}(\hat{e}_{gt}^2/\sum_{t=1}^{n}\hat{e}_{gt}^2 - \hat{e}_{ft}^2/\sum_{t=1}^{n}\hat{e}_{ft}^2)^2 \qquad (A.3)$$

since $\hat{\sigma}_s^2 = \sum_{t=1}^{n}\hat{e}_{st}^2/n$. As a result,

$$\hat{\omega}_n^2 \equiv \tfrac{1}{n}\sum_{t=1}^{n}\left[\log\frac{\mathcal{J}_f(y_t|\underset{\sim}{x}_{ft};\hat{\theta}_f)}{\mathcal{J}_y(y_t|\underset{\sim}{x}_{gt};\hat{\theta}_g)}\right]^2 - \left[\tfrac{1}{n}\sum_{t=1}^{n}\log\frac{\mathcal{J}_f(y_t|\underset{\sim}{x}_{ft};\hat{\theta}_f)}{\mathcal{J}_g(y_t|\underset{\sim}{x}_{gt};\hat{\theta}_g)}\right]^2$$

$$= \tfrac{n}{4}\sum_{t=1}^{n}(\hat{e}_{gt}^2/\sum_{t=1}^{n}\hat{e}_{gt}^2 - \hat{e}_{ft}^2/\sum_{t=1}^{n}\hat{e}_{ft}^2)^2. \qquad (A.4)$$

The proof is complete once we notice that $T_{fg} = n^{-1/2}LR_n(\hat{\theta}_f,\hat{\theta}_g)/\hat{\omega}_n$ and apply Vuong (1985, Theorem 5.2).

Proof of Lemma 4.2: (i) follows from Vuong (1985, Lemma 4.1) by noticing that the conditions of that lemma are satisfied under A2 - A5.

To prove (ii), note from (4.5) that H_o^{ω} is equivalent to:

$$[\underset{\sim}{x}'(\alpha^* - \gamma^*) + \underset{\sim}{z}'\beta^* - \underset{\sim}{w}'\delta^*][2y - \underset{\sim}{x}'(\alpha^* + \gamma^*) - \underset{\sim}{z}'\beta^* - \underset{\sim}{w}'\delta^*] = 0$$

H^o-almost surely. But given A2,

$\Pr\{(y,\underset{\sim}{x}): \ 2y = \underset{\sim}{x}'(\alpha^* + \gamma^*) + \underset{\sim}{z}'\beta^* + \underset{\sim}{w}'\delta^*\} = 0$. Thus

$\Pr\{(y,\underset{\sim}{x}): \ \underset{\sim}{x}'(\alpha^* - \gamma^*) + \underset{\sim}{z}'\beta^* - \underset{\sim}{w}'\delta^* = 0\} = 1$ which implies by A3 that $\beta^* = \delta^* = 0$ and $\alpha^* = \gamma^*$. To prove that (ii) implies H_o^{ω}, we note that when $\beta^* = 0$:

$$\begin{bmatrix} \sum_{xx}^o & \sum_{xz}^o \\ \sum_{zx}^o & \sum_{zz}^o \end{bmatrix} \begin{bmatrix} \alpha^* \\ 0 \end{bmatrix} = \begin{bmatrix} \sum_{xy}^o \\ \sum_{zy}^o \end{bmatrix} \qquad (A.5)$$

(see Lemma 2.1). Hence $\alpha^* = (\sum_{xx}^o)^{-1}\sum_{xy}^o$. Similarly, we can show that $\gamma^* = (\sum_{xx}^o)^{-1}\sum_{xy}^o$ when $\delta^* = 0$. Therefore $e_f = y - \underset{\sim}{x}'\alpha^* = y - \underset{\sim}{x}'\gamma^* = e_g$, H^o-almost surely.

To prove (iii), we show that (iii) and (ii) are equivalent. If $\beta^* = \delta^* = 0$, then $\alpha^* = (\sum_{xx}^o)^{-1}\sum_{xy}^o = \gamma^*$ and $\sum_{zx}^o\alpha^* = \sum_{zy}^o$ from (A.5). Similarly, $\sum_{wx}^o\alpha^* = \sum_{wy}^o$. It can be easily shown that $(\lambda_x,\lambda_z,\lambda_w) = (\alpha^*,0,0)$ is the unique solution for

$$\begin{bmatrix} \sum_{xx}^o & \sum_{xz}^o & \sum_{xw}^o \\ \sum_{zx}^o & \sum_{zz}^o & \sum_{zw}^o \\ \sum_{wx}^o & \sum_{wz}^o & \sum_{ww}^o \end{bmatrix} \begin{bmatrix} \lambda_x \\ \lambda_z \\ \lambda_w \end{bmatrix} = \begin{bmatrix} \sum_{xy}^o \\ \sum_{zy}^o \\ \sum_{wy}^o \end{bmatrix} \qquad (A.6)$$

Hence $(\lambda_x^*,\lambda_z^*,\lambda_w^*) = (\alpha^*,0,0)$ from Lemma 2.1. Conversely, if $\lambda_z^* = \lambda_w^* = 0$, then we can easily show that $(\alpha,\beta) = (\lambda_x^*,0)$ and $(\gamma,\delta) =$

$(\lambda_x^*, 0)$ are the unique solution for

$$\begin{bmatrix} \sum_{xx}^{o} & \sum_{xz}^{o} \\ \sum_{zx}^{o} & \sum_{zz}^{o} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{xy}^{o} \\ \sum_{zy}^{o} \end{bmatrix} \tag{A.7}$$

and

$$\begin{bmatrix} \sum_{xx}^{o} & \sum_{xw}^{o} \\ \sum_{wx}^{o} & \sum_{ww}^{o} \end{bmatrix} \begin{bmatrix} \gamma \\ \delta \end{bmatrix} = \begin{bmatrix} \sum_{xy}^{o} \\ \sum_{wy}^{o} \end{bmatrix} \tag{A.8}$$

respectively. Hence $\beta^* = \delta^* = 0$.

An example where $e_f^2 = e_g^2 \not\Rightarrow \beta^* = \delta^* = 0$: Consider the following two normal linear regression models for y given $\underline{x} = (z', w')'$ where z and w are both scalars:

$$M_f: \quad y = z\beta + \varepsilon_f, \quad \varepsilon_f \sim N(0, \sigma_f^2),$$

$$M_g: \quad y = w\delta + \varepsilon_g, \quad \varepsilon_g \sim N(0, \sigma_g^2).$$

Assume that the true joint p.d.f. of (y,z,w) is:

	w = -1			w = 1		
p.d.f	y = -6	y = 0	y = 6	y = -6	y = 0	y = 6
z = -1	2/12	0	1/12	0	1/4	0
z = 1	0	1/4	0	1/12	0	2/12

That is, $\Pr(y = -6, z = -1, w = -1) = 2/12$ and so forth. Then

$$\mathrm{Var}^o(y,z,w) = \begin{bmatrix} 18 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Hence, from Lemma 2.1, $\beta^* = \delta^* = 1$. But $\Pr(e_f^2 = e_g^2) = \Pr[(w - z)(2y - w - z) = 0] = 1$. In fact, it can easily be shown that A3-A4 are satisfied; the only violation is A2.

Proof of Lemma 4.3: Given (4.19) and (4.20), the first and second partial derivatives of $\log \mathcal{J}_s(y|\underline{x}_s;\theta_s)$ at $\theta_s = \theta_s^*$ under H_o^{ω} are:

$$\frac{\partial \log \mathcal{J}_s(y|\underline{x}_s;\theta_s^*)}{\partial \lambda_s} = \frac{\underline{x}_s e}{\sigma^{*2}}; \quad \frac{\partial \log \mathcal{J}_s(y|\underline{x}_s;\theta_s^*)}{\partial \sigma_s^2} = \frac{e^2}{2\sigma^{*4}} - \frac{1}{2\sigma^{*2}};$$

$$\frac{\partial \log^2 \mathcal{J}_s(y|\underline{x}_s;\theta_s^*)}{\partial \lambda_s \partial \lambda_s'} = -\frac{\underline{x}_s \underline{x}_s'}{\sigma^{*2}}; \quad \frac{\partial^2 \log \mathcal{J}_s(y|\underline{x}_s;\theta_s^*)}{\partial (\sigma_s^2)^2} = -\frac{e^2}{\sigma^{*6}} + \frac{1}{2\sigma^{*4}};$$

$$\frac{\partial^2 \log \mathcal{J}_s(y|\underline{x}_s;\theta_s^*)}{\partial \lambda_s \partial \sigma_s^2} = -\frac{\underline{x}_s e}{\sigma^{*4}}.$$

The lemma is then immediate upon replacing the above results into (4.16) - (4.18) and noting that $E^o(e^2) = \sigma^{*2}$ and $E^o(e\underline{x}_s) = 0$.

Proof of Lemma 4.4: Using the definition of W (i.e., equation (4.15)),

$$|W - \lambda I| = \begin{vmatrix} \frac{1}{\sigma^{*2}}C_f V_f - \lambda I & \frac{1}{\sigma^{*2}}U_f & \frac{1}{\sigma^{*2}}C_{fg}V_g & \frac{1}{\sigma^{*2}}U_f \\ \frac{1}{2\sigma^{*4}}U_f' V_f & \frac{m}{2\sigma^{*4}} - \lambda & \frac{1}{2\sigma^{*4}}U_g' V_g & \frac{m}{2\sigma^{*4}} \\ -\frac{1}{\sigma^{*2}}C_{fg}' V_f & -\frac{1}{\sigma^{*2}}U_g & -\frac{1}{\sigma^{*2}}C_g V_g - \lambda I & -\frac{1}{\sigma^{*2}}U_g \\ -\frac{1}{2\sigma^{*4}}U_f' V_f & -\frac{m}{2\sigma^{*4}} & -\frac{1}{2\sigma^{*4}}U_g' V_g & -\frac{m}{2\sigma^{*4}} - \lambda \end{vmatrix} \tag{A.9}$$

where

$$C_f = \begin{bmatrix} C_{xx} & C_{xz} \\ C_{zx} & C_{zz} \end{bmatrix}, \quad C_g = \begin{bmatrix} C_{xx} & C_{xw} \\ C_{wx} & C_{ww} \end{bmatrix}, \quad C_{fg} = \begin{bmatrix} C_{xx} & C_{xw} \\ C_{zx} & C_{zw} \end{bmatrix}$$

$V_s \equiv \sum^{o-1}_{\underline{x}_s \underline{x}_s}$, $U_s \equiv \mathrm{Cov}^o(\underline{x}_s e, e^2)$, $\forall s = f,g$; and $m = \mathrm{Var}^o(e^2)$. Now applying the following (block) row and (block) column operations:

(i) subtract column 4 from column 2,

(ii) add row 4 to row 2,

(iii) factorize λ from row 4 and column 4 respectively,

then $|W - \lambda I| = 0$ if and only if

$$\lambda^2 \begin{vmatrix} \frac{1}{\sigma^{*2}}C_f V_f - \lambda I & \frac{1}{\sigma^{*2}}C_{fg}V_g \\ -\frac{1}{\sigma^{*2}}C_{fg}' V_f & -\frac{1}{\sigma^{*2}}C_g V_g - \lambda I \end{vmatrix} = 0,$$

or alternatively,

$$\lambda^2 \begin{vmatrix} C_f - \lambda\sigma^{*2}\sum^{o}_{\underline{x}_f \underline{x}_f} & C_{fg} \\ C_{fg}' & C_g + \lambda\sigma^{*2}\sum^{o}_{\underline{x}_g \underline{x}_g} \end{vmatrix} = 0. \tag{A.10}$$

Upon substituting $\underline{x}_f' = (x', z')$, $\underline{x}_g' = (x', w')$ and applying the following (block) row and (block) column operations to (A.10):

(i) subtract column 1 from column 3,

(ii) interchange column 3 and column 4, row 3 and row 4 respectively,

(iii) factorize $(\lambda\sigma^{*2})^k$ from column 4,

(iv) subtract row 1 from row 4,

(v) factorize $(\lambda\sigma^{*2})^k$ from row 4,

then (A.10) implies

$$\det A \equiv \det \begin{bmatrix} C - \lambda\sigma^{*2} \begin{bmatrix} \sum^o_{xx} & \sum^o_{xz} & 0 \\ \sum^o_{zx} & \sum^o_{zz} & 0 \\ 0 & 0 & -\sum^o_{ww} \end{bmatrix} & \begin{matrix} \sum^o_{xx} \\ \sum^o_{zx} \\ \sum^o_{wx} \end{matrix} \\ \begin{matrix} \sum^o_{xx} & \sum^o_{xz} & \sum^o_{xw} \end{matrix} & 0 \end{bmatrix} = 0.$$

Postmultiplying by the non-singular matrix

$$B \equiv \begin{bmatrix} I_k & -\sum_{xx}^{o-1}\sum_{xz}^{o} & -\sum_{xx}^{o-1}\sum_{xw}^{o} & 0 \\ 0 & I_p & 0 & 0 \\ 0 & 0 & I_q & 0 \\ 0 & 0 & 0 & I_k \end{bmatrix}$$

and premultiplying by B', we obtain:

$$B'AB = \begin{bmatrix} C_{xx} - \lambda\sigma^{*2}\sum_{xx}^{o} & ; & C_{x\cdot}R + \lambda\sigma^{*2}(0, \sum_{xw}^{o}) & ; & \sum_{xx}^{o} \\ R'C_{\cdot x} + \lambda\sigma^{*2}\begin{bmatrix} 0 \\ \sum_{wx}^{o} \end{bmatrix} & ; & R'CR - \lambda\sigma^{*2}\mathrm{Diag}(Q_{zz}, -Q_{ww}) & ; & 0 \\ \sum_{xx}^{o} & ; & 0 & ; & 0 \end{bmatrix}$$

where R is given by (4.24) and $C_{x\cdot} = C_{\cdot x}'$ is the first row block of C. Interchanging the first column block and the third column block, the matrix $B'AB$ becomes block upper-triangular. Since $\sum_{xx}^{o}$ is non-singular, then $\det B'AB = 0$ is equivalent to

$$\det\left[R'\left[C - \lambda\sigma^{*2} \begin{bmatrix} \sum_{xx}^{o} & \sum_{xz}^{o} & 0 \\ \sum_{zx}^{oo} & \sum_{zz}^{o} & 0 \\ 0 & 0 & -\sum_{ww}^{o} \end{bmatrix} \right] R \right] = 0.$$

Combining the above results, and calculating the coefficient matrix associated with $\lambda\sigma^{*2}$, $|W - \lambda I| = 0$ if and only if

$$\lambda^{2k+2}\sigma^{*4k} \cdot \det[R'CR - \lambda\sigma^{*2}\mathrm{Diag}(Q_{zz}, -Q_{ww})] = 0. \tag{A.11}$$

The proof is now complete.

Proof of Proposition 4.5: Immediate from Vuong (1985, Theorem 4.3).

Proof of Lemma 4.6: Part (i) follows from (3.3), (4.24), (4.28), and the partitioned inverse formula applied successively to $\hat{\underline{\lambda}}_f = (\hat{\alpha}', \hat{\beta}')'$ and $\hat{\underline{\lambda}}_g = (\hat{\gamma}', \hat{\delta}')'$.

To prove (ii), we note that under H_o^{ω}:

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \underline{x}_t e_t \xrightarrow{D} N(0, C) \tag{A.12}$$

by the multivariate central limit theorem, where C is given by (4.25) and $e_t = y_t - x_t'\alpha^* = y_t - x_t'\gamma^*$. Moreover, we have:

$$\hat{\sum}_{\underline{x}y} = \hat{\sum}_{\underline{x}x}\alpha^* + \frac{1}{n}\sum_{t=1}^{n} \underline{x}_t e_t. \tag{A.13}$$

Thus (4.34) becomes

$$(\hat{\beta}', \hat{\delta}')' = \mathrm{Diag}(\hat{Q}_{zz}^{-1}, \hat{Q}_{ww}^{-1})\hat{R}'[\hat{\sum}_{\underline{x}x}\alpha^* + \frac{1}{n}\sum_{t=1}^{n} \underline{x}_t e_t]$$

$$= \mathrm{Diag}(\hat{Q}_{zz}^{-1}, \hat{Q}_{ww}^{-1})\hat{R}'\frac{1}{n}\sum_{t=1}^{n} \underline{x}_t e_t \tag{A.14}$$

where the second equation follows from $\hat{R}'\hat{\sum}_{\underline{x}x} = 0$. Part (ii) follows by applying (A.12).

Proof of Proposition 4.7: Immediate from Vuong (1986, Theorems 1 and 2).

Proof of Proposition 4.8: Part (i) follows from (3.3), (4.24), (4.28), and the partitioned inverse formula applied to $\hat{\underline{\lambda}} = (\hat{\underline{\lambda}}_x', \hat{\underline{\lambda}}_z', \hat{\underline{\lambda}}_w')'$.

Since (A.13) holds under H_o^ω, then (4.38) becomes:

$$(\hat{\underline{\lambda}}_z', \hat{\underline{\lambda}}_w')' = \hat{Q}^{-1}\hat{R}'\frac{1}{n}\sum_{t=1}^{n}\underline{x}_t e_t \tag{A.15}$$

where we have used $\hat{R}'\hat{\sum}_{\underline{x}\underline{x}} = 0$. Part (ii) follows by applying (A.12).

Proof of Proposition 4.9: Because $\text{Diag}(Q_{zz}, Q_{ww})$ is non-singular, it can easily be shown that a matrix G is a g-inverse of V if and only if it is of the form

$$G = \text{Diag}(Q_{zz}, Q_{ww})[R'CR]^{-}\text{Diag}(Q_{zz}, Q_{ww}), \tag{A.16}$$

where $[R'CR]^{-}$ is a g-inverse of $R'CR$. Similarly, since Q is non-singular, a matrix H is a g-inverse of W if and only if it is of the form

$$H = Q[R'CR]^{-}Q. \tag{A.17}$$

Choose now the same g-inverse $[R'CR]^{-}$ in (A.16) and (A.17). Then, let $G_n = \text{Diag}(\hat{Q}_{zz}, \hat{Q}_{ww})K_n\text{Diag}(\hat{Q}_{zz}, Q_{ww})$ and $H_n = \hat{Q}K_n\hat{Q}$ where $K_n \xrightarrow{a.s.} (R'CR)^{-}$. For this choice of g-inverses, we have, using (4.34), (4.36), (4.38), and (4.40):

$$W_n^1 = W_n^2 = n\hat{\sum}_{y\underline{x}}\hat{R}K_n\hat{R}'\hat{\sum}_{\underline{x}y}. \tag{A.18}$$

Part (i) follows from Vuong (1986, Theorem 1), since under H_o^ω, the W_n^1's are equivalent for any choice of G, while the W_n^2's are equivalent for any choice of H.

Part (ii) is immediate from (i) and Proposition 4.8 - (i). Part (iii) is immediate from Vuong (1986, Theorem 2).

Proof of Lemma 5.1: Given A6 and Corollary 2.2, we have $\underline{\lambda}^o = \underline{\lambda}^*$ and $\sigma_o^2 = \sigma_{f\vee g}^{*2}$. The result follows from Lemma 4.2, (4.12) and (4.13).

Proof of Lemma 5.2: Upon substituting (5.3) into (4.30), we have

$$\det\begin{bmatrix}(1-\lambda_\omega)Q_{zz} & Q_{zw}\\ Q_{wz} & (1+\lambda_\omega)Q_{ww}\end{bmatrix} = 0 \tag{A.19}$$

For any $\lambda_\omega \neq 1$, the above equation reduces to

$$|(1-\lambda_\omega)Q_{zz}||Q_{wz}(1-\lambda_\omega)^{-1}Q_{zz}^{-1}Q_{zw} - (1+\lambda_\omega)Q_{ww}| = 0,$$

i.e., $|Q_{wz}Q_{zz}^{-1}Q_{zw} - (1-\lambda_\omega^2)Q_{ww}| = 0$, which is (5.4). But (A.19) has p + q solutions, while (5.4) only has 2q solutions, hence the other p - q eigenvalues must be one. Moreover, from (5.4), the number of solutions $\lambda_\omega^2 = 1$ which satisfy it is the same as the number of zero eigenvalues for $Q_{ww}^{-1/2}Q_{wz}Q_{zz}^{-1}Q_{zw}Q_{ww}^{-1/2}$, which is in turn equal to rank Q_{wz}. Hence totally we have p - q + (q - rank Q_{wz}) = p - rank Q_{wz} one eigenvalues and q - rank Q_{wz} minus one eigenvalues. We now only have

to show that $0 < \lambda_\omega^2 \leq 1$. First, since $(1 - \lambda_\omega^2)$ is the eigenvalue of a p.s.d. matrix $Q_{ww}^{-1/2}Q_{wz}Q_{zz}^{-1}Q_{zw}Q_{ww}^{-1/2}$, hence $\lambda_\omega^2 \leq 1$. Furthermore,

$$0 \neq \det\begin{bmatrix} \sum_{xx}^{o} & \sum_{xz}^{o} & \sum_{xw}^{o} \\ \sum_{zx}^{o} & \sum_{zz}^{o} & \sum_{zw}^{o} \\ \sum_{wx}^{o} & \sum_{wz}^{o} & \sum_{ww}^{o} \end{bmatrix} = |\sum_{xx}^{o}|\det\begin{bmatrix} Q_{zz} & Q_{zw} \\ Q_{wz} & Q_{ww} \end{bmatrix}$$

$$= |\sum_{xx}||Q_{zz}||Q_{ww} - Q_{wz}Q_{zz}^{-1}Q_{zw}|$$

Therefore $\lambda_\omega = 0$ is not a solution for (5.4). The proof is now complete by using Lemma 4.4.

Proof of Proposition 5.3: Immediate from Proposition 4.5, Lemma 5.2, and the definition of a weighted sum of chi-squares.

Proof of Corollary 5.4: Obvious.

Proof of Proposition 5.5: The numerical equivalences between W_n^1 and W_n^2 follows from (A.18) since g-inverses need not be used. Moreover, $K_n = (\hat{R}'\hat{C}\hat{R})^{-1} = \hat{\sigma}^{-2}(\hat{R}'\hat{\sum}\hat{R})^{-1}$. But $\hat{R}'\hat{\sum}\hat{R} = \hat{Q}$. This establishes (i). Parts (ii) and (iii) follow from Propositions 4.7 and 4.9.

Proof of Proposition 6.1: Immediate from Lemma 5.2 and Vuong (1985, Theorem 6.4).

Proof of Corollary 6.2: Obvious.

FOOTNOTES

* This research was partially supported by National Science Foundation Grant SES-8410593. A preliminary draft of this paper was presented at the Southern California Econometric Conference at Lake Arrowhead, 1986. We are grateful to A. Golberger for helpful remarks and to D. Rivers for expected comments. The second author also thanks S. Heart for stimulating thoughts.

1. The observations $(y_t, \underline{x}_t')'$ can be obtained, for instance, by random sampling from a population with joint distribution H^o. Alternatively, they can be obtained by stratified exogenous sampling, in which case H^o is the product of the conditional distribution of y given $\underline{x}$ times the marginal distribution of $\underline{x}$.
2. In general, θ^+ is not unique. See below and footnote 5.
3. Given our subsequent use of the KLIC, we can assume without loss of generality that $E^o[\log h^o(y|\underline{x})]$ is finite.
4. For instance, unlike the KLIC, the MSE criterion does not distinguish between two competing models having identical parametric specification for the conditional mean of y given $\underline{x}$.
5. The value σ_s^{+2} is not unique since the MSE-distance (2.3) does not depend on σ_s^2 when the model is a normal linear regression.
6. These authors have derived the asymptotic distribution of the statistic $\Delta\sigma^2$ under the additional assumption that $\phi_f(\cdot|\cdot;\theta_f^*) \neq \phi_g(\cdot|\cdot;\theta_g^*)$ H^o-almost surely, where $\phi_s(\cdot|\cdot;\theta_s)$ denotes the univariate normal density for y given $\underline{x}_s$ with parameters θ_s. As emphasized in the next sections, the fact that this assumption

may be violated much complicates the analysis.

7. The idea of discriminating between M_f and M_g by testing $\sigma_g^{*2} = \sigma_f^{*2}$ dates back to Hotelling (1940) who proposed, for the single explanatory variable case, a test of the more restrictive hypothesis that the correlation coefficients $\rho_{y\underline{x}_f} = \rho_{y\underline{x}_g}$ under the additional assumption that the true conditional distribution of y given $(\underline{x}_f, \underline{x}_g)$ is normal with linear conditional mean and constant conditional variance. For a generalization to more than one explanatory variables, see Chow (1980).

8. It can be shown that $\Delta\sigma^2$, ΔC_p, and ΔPC are biased estimators of $\Delta MSE(M_f, M_g) = \sigma_g^{*2} - \sigma_f^{*2}$ even when M_f and M_g are both correctly specified. On the other hand, it can be shown that the statistic $[\hat{\sigma}_g^2 + \hat{\sigma}^2 \ell_g/(n - \ell)] - [\hat{\sigma}_f^2 + \hat{\sigma}^2 \ell_f/(n - \ell)]$ and the statistic $[n\hat{\sigma}_g^2/(n - \ell_g)] - [n\hat{\sigma}_f^2/(n - \ell_f)]$ are both unbiased estimator of $\sigma_g^{*2} - \sigma_f^{*2}$ for fixed explanatory variables, the former under the assumption that $M_{f \vee g}$ is correctly specified, and the latter under the assumption that M_f and M_g are correctly specified.

9. For a generalization of Sawa criterion to non-linear models, see Chow (1981).

10. As the proof shows, the importance of Assumption A2 is to ensure that $e_f^2 = e_g^2$ H^o-almost surely is equivalent to $e_f = e_g$ H^o-almost surely.

11. To ensure that these matrices exist, we assume, in addition to A3, that all fourth moments of the vector $(y, \underline{x}')'$ exist.

12. The matrix W simplifies if the competing models are asymptotically orthogonal, i.e., if $B_{fg} = 0$. Unfortunately, it can be shown that normal linear regression models can never be asymptotically orthogonal.

13. The exact number of zero eigenvalues of W is $2k + 2 + [p + q\text{-rank } R'CR]$. The non-zero eigenvalues of W are the non-zero eigenvalues of $(\sigma^*)^{-2} R'CR \text{ Diag}(Q_{zz}^{-1}, -Q_{ww}^{-1})$.

14. This ensures that, for any choice of g-inverse G and any consistent sequence $\{G_n\}$, the statistic W_n^1 is asymptotically chi-square distributed under the null hypothesis $\beta^* = \delta^* = 0$ (see Vuong (1986)). An alternative and more frequent method is to use $(V_n)^-$ in place of G_n where V_n is a consistent estimator of V. Additional conditions on V_n must, however, be imposed to insure the limiting chi-square distribution (see Andrews (1986) for the difficulties associated with this latter method.)

15. From Vuong (1986), the Wald statistics (4.36) based on different choice of g-inverses G are asymptotically equivalent under H_o^{ω}.

16. As the proof of Proposition 4.9 shows, for some particular choices of G, H, G_n, and H_n, the Wald statistics W_n^1 and W_n^2 are numerically equal.

17. While $Q_{zw} = 0$ greatly simplifies the variance test and the LR test for model selection discussed in this and the next sections, this condition is precisely the one under which the Cox type tests cannot be applied (see, e.g., Pesaran (1974, p.158), Pesaran and Deaton (1978, p. 681), Davidson and MacKinnon (1981,

p.785), and White (1982, p.318)).

18. If the larger model is correctly specified, $\beta^* = \beta^o$ and we have the classical hypothesis testing under correct specification.

19. To test H_o^θ against H_A^θ, Vuong (1985, Theorem 7.4) considers twice the LR statistic which is in general asymptotically distributed as a weighted sum of chi-squares. One can also consider White (1982a)'s robust Wald and LM statistics. Asymptotic comparison of the tests based on these statistics is left for future research.

20. For a survey see Mackinnon (1983). Other non-nested hypothesis tests are the J and P tests proposed by Davidson and MacKinnon (1981) which are in fact Cox-type tests using a simpler estimator of $\Delta KLIC(F_\theta, G_\gamma)$ than the one initially proposed by Cox (see White (1982b)).

REFERENCES

Akaike, H. "Information Theory and an Extension of the Likelihood Ratio Principle." In Proceedings of the Second International Symposium of Information Theory, edited by B. N. Petrov and F. Csaki, 257-281. Budapest: Akademiai Kiodo, 1973.

Amemiya, T. "Selection of Regressors." International Economic Review 21 (1980):331-354.

Andrews, D. W. K. "Asymptotic Results for General Wald Tests." Cowles Foundation Discussion Paper No. 761, 1985.

Atkinson, A. C. "A Method for Discriminating between Models." Journal of the Royal Statistical Society, Series B, 32 (1970):323-353.

Chow, G. C. "The Selection of Variables for Use in Prediction: A Generalization of Hotelling's Solution." In Quantitative Econometrics and Development, edited by L. N. Klein, M. Nerlove, and S. C. Tsiang, 105-114. New York: Academic, 1980.

_____. "Selection of Econometric Models by the Information Criterion." In Proceedings of Econometrica Society European Meeting, edited by E. G. Charatsis. Amersterdam: North Holland, 1981.

_____. Econometrics. New York: McGraw-Hill, 1983.

Cox, D. R. "Tests of Separate Families of Hypotheses." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1 (1961):105-123.

_____. "Further Results on Tests of Separate Families of Hypotheses." Journal of the Royal Statistical Society, Series B, 24 (1962):406-424.

Dastoor, N. K. "A Note on the Interpretation of the Cox Procedure for Non-Nested Hypotheses." Economics Letters 8(1981):113-119.

Davidson R. and MacKinnon, J. G. "Several Tests for Model Specification in the Presence of Alternative Hypotheses." Econometrica 49 (1981):781-793.

Gaver, K. M. and Geisel, M. S. "Discriminating Among Alternative Models: Bayesian and Non-Bayesian Methods." In Frontiers in Econometrics, edited by P. Zarembka, 48-80. New York: Academic, 1974.

Hocking, R. R. "The Analysis and Selection of Variables in Linear Regression." Biometrics 32 (1976):1-49.

Hotelling, H. "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters." Annals of Mathematical Statistics 11 (1940):271-283.

Johnson, N. L. and Kotz, S. Distribution in Statistics: Continuous Multivariate Distributions. New York: John Wiley and Sons, 1972.

Judge, G. G., et. al. The Theory and Practice of Econometrics. New York: John Wiley and Sons, 1985.

Kullback, S. and Leibler, R. A. "On Information and Sufficiency." Annals of Mathematical Statistics 22 (1951):79-86.

Leamer, E. "Information Criterion for Choice of Regression Models: A Comment." Econometrica 47 (1979):507-510.

Lindley, D. V. "The Choice of Variables in Multiple Regression." Journal of the Royal Statistical Society, Series B, 30 (1968):31-66.

MacKinnon, J. G. "Model Specification Tests Against Non-Nested Alternatives." Econometric Reviews 2 (1983):85-110.

Mallows, C. L. "Some Comment on C_p." Technometrics 15 (1973):661-675.

Moore, D. S. "Generalized Inverses, Wald's Method, and the Construction of Chi-squared Tests of Fit." Journal of the American Statistical Association 72 (1977):131-137.

Pesaran, M. H. "On the General Problem of Model Selection." Review of Economic Studies 41 (1974):153-171.

_____. "On the Comprehensive Method of Testing Non-Nested Regression Models." Journal of Econometrics 18 (1982):263-274.

Pesaran, M. H. and Deaton, A. S. "Testing Non-Nested Nonlinear Regression Models." Econometrica 46 (1978):677-684.

Sawa, T. "Information Criteria for Discriminating Among Alternative Regression Models." Econometrica 46 (1978):1273-1291.

Schwarz, G. "Estimating the Dimension of a Model." Annals of Statistics 6 (1978):461-464.

Vuong, Q. H. "Misspecification and Conditional Maximum Likelihood Estimation." Social Science Working Paper no. 503. Pasadena: California Institute of Technology, 1983.

_____. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." Mimeo. Pasadena: California Institute of Technology, 1985.

_____. "Generalized Inverses and Asymptotic Properties of Wald Tests." Mimeo, 1986.

White, H. "Consequences and Detection of Misspecified Nonlinear Regression Models. Journal of the American Statistical Association 76 (1981):419-433.

_____. "Maximum Likelihood Estimation of Misspecified Models." Econometrica 50 (1982a):1-25.

_____. "Regularity Conditions for Cox's Test of Non-nested Hypotheses." Journal of Econometrics 19 (1982b):301-318.

White, H. and Olson, L. "Determinants of Wage Change on the Job: A Symmetric Test of Non-Nested Hypotheses." Mimeo. Rochester: University of Rochester, 1979.

Zellner, A. An Introduction to Bayesian Inference in Econometrics. New York: Wiley, 1971.